Investigating singing voice: quantitative and qualitative approaches to studying cross-cultural vocal production

Polina Proutskova

PhD Thesis

Department of Computing
Goldsmiths, University of London

6 October 2017

I hereby declare that, except where explicit attribution is made, the work presented in this thesis is entirely my own.

PolinaProutskova

Abstract

This thesis was motivated by an experiment carried out in the 1960s that studied the relationship between vocal performance practice and society by means of statistical analysis. Using a comprehensive corpus of audio recordings of singing from around the world collected over several decades, the ethnomusicologist Alan Lomax devised the Cantometrics project, the largest comparative study of music, in which 36 performance practice characteristics were rated for each recording. With particular interest in vocal production, we intended to formalise the knowledge of vocal production to enable statistical and computational approaches in the spirit of Cantometrics.

Three models of vocal production were investigated: the perceptual model from Cantometrics, a physical model from voice science and a physiological model from singing education. We built on Johan Sundberg's vocal source parameters and Jo Estill's physiological building blocks as the basis to develop an ontology of vocal production.

Two approaches to automated characterisation of the ontological descriptors were considered. For the incremental approach a proof-of-concept experiment on automatic labelling of phonation modes was presented, based on reconstructing the vocal source waveform by means of inverse filtering. We created a dataset of sustained sung vowels with annotations on pitch, vowel and phonation mode on which our model was trained. Steps to generalise this experiment to more complex data were outlined, discussing the challenges of such generalisation.

The integrated approach addressed the full variance in the data, turning to the methodology of expert knowledge elicitation in order to annotate the original Cantometrics dataset with our descriptors. We performed an investigative mixed-methods study in which 13 vocal physiology experts from different professional backgrounds were interviewed; they used our ontology to analyse vocal production in the Cantometrics dataset. The goal of the study was to: a) validate the acceptance of our ontological terms, b) verify the consensus between experts on the values of the descriptors, c) collect reliable annotations. While the acceptance of the ontology

was good for most terms, quantitative analysis showed good agreement between experts for only two out of 11 descriptors (*larynx height*, *aryepiglottic sphincter*). A detailed qualitative analysis of the interview data (over 33 hours) was followed by a meta-analysis extracting common themes and confounding issues which point to probable reasons for the disagreement. For *aryepiglottic sphincter* and *larynx height* we collected the average ratings, which constitute the first set of reliable annotations on vocal production. A strong correlation was found between *larynx height* and the *vocal width* parameter from Cantometrics; *larynx height* was therefore a good candidate to replace *vocal width* as a more objective descriptor.

The current work was based on knowledge from a number of research disciplines, and its results are discussed from the viewpoint of several fields – MIR, vocal pedagogy, Cantometrics – for which they present significant implications. Future research is suggested for each of the fields. Based on the meta-analysis, we account for the reasons for disagreement between experts on the subject of vocal production, from music information retrieval (MIR) and singing education perspectives. We further explain the various kinds of bias that affect raters.

We conclude that vocal physiology, though offering a more objective language than perceptual descriptors, is not well-suited as an ontological middle layer for statistical approaches to singing given the current state of knowledge. A mixed perceptual-objective path to ontology building is suggested and ways to collect reliable annotations are outlined.

In the domain of vocal pedagogy we touch on the issue of communication on vocal physiology between experts, between teacher and student; we consider the future of teaching vocal technique and make suggestions for new experiments in the field.

A plan is presented for revising and scaling up Cantometrics as an interdisciplinary collaboration. Possible contributions of MIR, ethnomusicologists and vocal production specialists are specified.

Acknowledgements

I would first like to thank my supervisors Prof. Geraint Wiggins, Dr Christophe Rhodes and Prof. Tim Crawford for their extraordinary support in this thesis process. The intellectual discussion, encouragement and guidance and the critique they offered have been invaluable. I am particularly grateful to Prof. Wiggins for taking over my supervision after a difficult period and helping me to get back on track, ultimately leading to the thesis completion. My thanks go to Dr Rhodes for helping me to navigate through the bureaucratic maze.

My thanks to Prof. Michael Casey for introducing me to music informatics and to Prof. Mark D'Inverno for his help. In addition, I would like to thank Lesley Hewings and the Graduate School for supporting my case.

I would like to sincerely thank Dr Victor Grauer, the co-inventor of Cantometrics, for suggesting to focus on the subordination of women hypothesis, which gave the initial direction for the current work. His explorations on the evolution of musical style have been a great inspiration that motivated me to return to academic research.

My special thanks go to Anna Lomax Wood, the Alan Lomax Archives and the Association for Cultural Equity for interest in my research and for providing me with the recordings from the Cantometrics Training Tapes which have been the subject of my research ever since.

I am very grateful to Prof. Johan Sundberg for his involvement with my recordings for the Phonation Modes Dataset and his general support. His Summer School taught me more about singing voice in two weeks than I had learned in years.

The participants in my study require a particular mention: I greatly appreciate that in spite of all their other commitments they found time to take part and shared their knowledge and expertise with me with openness, intellectual rigour and trust.

I am heartily thankful to Dr Gillyanne Kayes for her support and advice and for being a role model. My sincere thanks to Kim Chandler for introducing me to the British Voice Association and to the science minded singing voice community.

I am grateful to Dr Daniel Müllensiefen, Dr Kat Agres and Vincent Akkermans for their valuable recommendations.

I'd like to thank my singing teacher Natalia Vladimirovna Silina for her steadfast guidance and support through all these years and her dedication to vocal pedagogy that inspired my love of studying the voice.

My husband Dr Sven Macholl has been my greatest supporter through all the long years this PhD took to mature and to come to completion. Without his care, understanding and patience it would not have happened. I am indebted by my family for believing in me, giving me space and time to work and cheering me up whenever I felt lost.

My gratitude includes my Mum, who sacrificed her own PhD in order to raise me. It is the values she instilled in me – the love of learning, the importance of knowledge and the respect for the academic world – that have led me to where I am today. My mother's kindness has always been my biggest resource. I am so happy that I can share this achievement with her.

Last but not least, my son has grown up alongside my PhD, filling my days with joy and meaning. Seeing how mature and self-reliant he has become makes clear what a long way it took for this body of work to shape into the thesis presented here.

Contents

List of Tables

List of Figures

Main contributions

Main contributions to knowledge

This PhD was concerned with formalising the language about vocal production in singing to the extent that it can be used in computational modelling. This was achieved by compiling an ontology of vocal production (Chapter 2) and verifying experts' acceptance of the ontological terms (Section 5.1). It was demonstrated that a subjective and inconsistent descriptor of vocal production – *vocal width* from the Cantometrics system – can be mapped onto more objective descriptors from the complied ontology (Section 5.8).

It was also shown that for existing recordings of singing objective annotation with the ontological terms is problematic. Experts annotations were found to be consistent for only two out of eleven descriptors (Chapter 5). An extensive qualitative analysis of confounding issues and reasons for experts' disagreement was conducted (Chapter 6), leading to recommendations on bias in annotations (Chapters 7 and 8). Further contributions to knowledge include:

- Created and published the Phonation Modes Dataset: a dataset of recorded audio of sung vowels, produced under studio conditions, sung in multiple phonation modes on multiple pitches (Section 3.2, Appendix Section 8.3.3).
- Implemented in R language Prof. Krippendorff's bootstrapping algorithm for computing confidence intervals for the Krippendorff's alpha statistic and shared it as open source (Section 5.4), extending its functionality to weighted observations (Section 5.7).
- Published the first curated cross-cultural dataset with reliable annotations on vocal production (Section 5.8).
- Analysed the results from the viewpoint of different disciplines: MIR (music information research), vocal pedagogy, as well as Cantometrics; made future research suggestions for each of the fields (Chapter 8).

Cantometrics

This thesis offers a vision of revising the Cantometrics experiment employing modern technological approaches, overcoming some of its methodological issues and widening its scope to millions of recordings.

To begin, Cantometrics descriptors related to vocal production were scrutinised, in particular the *vocal width* parameter (Section 2.3); hypotheses about the contributions of various physiological phenomena to the perception of *vocal width* were put forward (Section 2.4).

Then more objective descriptors of vocal production were identified for which reliable ratings can be produced; these ratings were collected for 11 Cantometrics tracks (Section 5.8). Two physiological descriptors (*larynx height*, *AES*) were found to correlate with the Cantometrics ratings of *vocal width*.

The thesis suggests a new objective-subjective approach to ontology building for vocal production and a roadmap of Cantometrics revision in this mode (Section 8.1).

MIR

In terms of MIR this PhD is about the main barrier for new MIR research – the lack of datasets with reliable annotations (ground truth). It suggests a methodology (Chapter 4) and conducts a proof-of-concept experiment (Chapters 5 and 6) for an understudied field of vocal production where the state of knowledge is not sufficient to allow predictions and direct measurements are generally not available. It addresses the questions: why there are no reliable annotations (Chapter 1), how to generate new annotations from expert knowledge (Chapter 4), under what conditions they will be reliable (Section 4.1), how to elicit the main confounding issues and reasons for disagreement between experts (Chapter 7).

A position is taken concluding from the research that vocal physiology is not well suited as a model or a middle layer for automatic approaches to singing, given the current state of knowledge. Future prospects are discussed (Section 8.2).

Vocal pedagogy

From the viewpoint of singing education this thesis investigates quantitatively (Chapter 5) and qualitatively (Chapter 6) how experts deduce physiological settings and processes in singing through auditory-perceptual analysis; it uncovers the large extent to which they disagree in their analysis; elicits the main reasons for

disagreement and the various kinds of bias they are subject to (Chapter 7). Further experiments to investigate bias and disagreement are proposed (Section 8.3).

Future research suggestions

- further evaluation of collected data: investigating confidence of ratings, salience of descriptors and perceptual ratings by our participants (Section 8.1.1)
- a mixed objective-perceptual approach to ontology building for a Cantometrics revision and for studying singing in general (Section 8.1.3)
- devising a training system for raters of singing that would help to ground their ratings and equalise the perceptual singing spaces (mental representations of singing) for various raters (Section 8.1.6.1)
- an online game/app for collecting information on perceptual descriptors of singing and verifying their universality (Sections 8.1.6.2 and 8.2.4.1)
- a roadmap to revising and scaling up Cantometrics based on the mixed objective-perceptual approach (Section 8.1.6.5), detailing the contribution by MIR researchers (Section 8.2.4) and singing voice professionals (Section 8.3.3)
- continuing work on the Phonation Modes Dataset: conduct an independent evaluation of annotations; add further recordings by other singers (Section 8.2.5)
- investigate one of MIR's basic assumptions: that results obtained for audio datasets would hold for other contexts, e.g. those including a visual aspect (Section 8.2.5)
- explore possible benefits of employing vocal physiology for singing voice recommendations as well as genre classification (Section 8.2.5)
- proposed an experiment on experts' consensus about vocal physiology that would eliminate cultural bias (Section 5.10)
- consider changing the granularity/time scale in future studies to account for vowel changes (Section 8.2.4.3)

experiment with empathic listening: e.g. compare experts' verbal reports/reflections on the analysis process and stroboscopic pictures (Section 8.3.3)

Author's relevant publications

Publications from this thesis

2016 [1] **2016** (with Christophe Rhodes, Tim Crawford and Geraint Wiggins). "Formalising Cross-Cultural Vocal Production". In: *6th International Workshop of Folk Music Analysis - FMA, proc.*

2015 [2] **2015** (with Christophe Rhodes, Tim Crawford and Geraint Wiggins). "Approaching Vocal Production In World's Music Cultures – A Mixed Methods Study Based On The Physiology Of Singing". In: *5th International Workshop of Folk Music Analysis - FMA, proc.* Pp. 95–99.

2014 [3] **2014** (with Christophe Rhodes, Tim Crawford and Geraint Wiggins). "Ontological description of vocal production in world's music cultures–a physiological approach". In: *International Conference of Students of Systematic Musicology*.

[4] **2014** (with Geraint Wiggins, Christophe Rhodes and Tim Crawford). "Vocal production in world's music cultures". In: *Third International Conference on Analytical Approaches to World Music, Proc.*

2013 [5] **2013**. "MIR model of vocal timbre in world's cultures - where do we start". In: *III International Workshop of Folk Music Analysis - FMA, proc.* Pp. 93–94.

[6] **2013** (with Christophe Rhodes, Tim Crawford and Geraint Wiggins). "Breathy, Resonant, Pressed - Automatic Detection Of Phonation Mode From Audio Recordings of Singing". In: *Journal of New Music Research* 42.2.

2012 [7] **Apr. 2012**. "Does singing style correlate to social behaviour? - A revision of the Cantometric descriptor vocal tension and its correlation to the subordination of women in society". In: *Boundaries between Genres: Flamenco and Other Musical Oral Traditions III Interdisciplinary Conference on Flamenco Research - INFLA II International Workshop of Folk Music Analysis - FMA*. Ed. by José Miguel Díaz Báñez, Francisco Javier Escobar Borrego, and Inmaculada Ventura Molina. Sevilla: Escuela Técnica Superior de Ingeniería, pp. 235–238.

[8] **2012** (with Christophe Rhodes, Geraint Wiggins and Tim Crawford). "Breathy Or Resonant – A Controlled And Curated Dataset For Phonation Mode Detection In Singing". In: *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*.

Code and datasets

2017 [9] **Aug. 2017** (with Mike Gruszczynski). *kripp.boot - An R Package for Performing Bootstrap Replicates of Krippendorff's alpha on Intercoder Reliability Data*. URL: `https://github.com/MikeGruz/kripp.boot`.

2016 [10] **Nov. 2016**. *Phonation Modes Dataset*. https://osf.io/pa3ha/.

[11] **2016b**. *Vocal Production Ontology*. https://osf.io/pff8m/.

Other relevant publications by the author

2013 [12] **2013** (with Michela Magas). "A location-tracking interface for ethnomusicolog-

ical collections". In: *Journal of New Music Research* 42.2.

2012 [13] **Nov. 2012**. *Digital libraries for music and voice*. Tech. rep. AIRS project. URL: `http://www.airsplace.ca/node/1124`.

2009 [14] **2009** (with Michael Casey). "You call THAT singing? Ensemble classification for multi-cultural collections of music recordings". In: *Proceedings of the International Symposium on Music Information Retrieval*.

2008 [15] **July 2008**. "Data infrastructure for ethnomusicological archives - current situation and future perspectives". In: *Journal of International Society of Sound- and Audiovisual Archives* 31, pp. 45–54.

2007 [16] **2007**. "Musical Memory of the World - Data Infrastructure in Ethnomusicological Archives". In: *Proceedings of the International Symposium on Music Information Retrieval*.

2006 [17] **2006** (with Mirjam Gericke, Maurice Mengel and Deniza Popova). "Diversity of musical expression in Europe". In: *The Protection and Promotion of Musical Diversity*. Ed. by Richard Letts. UNESCO, International Music Council, pp. 308–411.

1 Introduction

A phenomenon of singing is present in every human culture and is often highly valued: while in some places being an accomplished singer is reflected in a higher status or wealth, anywhere in the world a good singer attracts love and admiration. In Western culture, the popular music industry generates a formidable revenue through what is mainly vocal music; in the US 23.5 million people were involved in choral performances according to the 2004 report (Bell 2004) – by far more than in any other form of artistic expression. We love singing, we love listening to good singing – it is one of the worlds's wonders that is available to every human being and seems to possess such an overwhelming power over us. As some say, it is something worth living for.

Yet there is a lot of mystery around it – the scientific understanding of singing remains limited. Researchers debate about its origins, whether it predated the development of language, and what kind of evolutionary advantage, if any, it might have presented to humans (Pinker 1997; Lieberman 1998; Trevarthen 1999-2000; Wallin, Merker and Brown 2000; Miller 2000; Merker 2000; Huron 2001; Mithen and Bannan 2004; Mithen 2005; Cross 2006; Dunbar 2012; Cross 2012; Pearce, Launay and Dunbar 2015). While systematic teaching of singing in the Western culture has been documented since the 13th century, the study of its mechanisms – the physiology – only began in the 19th century with Manuel Garcia's investigations of the functioning glottis by means of a laryngoscope. There remain numerous gaps in our knowledge of vocal physiology, and voice science has yet to develop a comprehensive model of the singing voice. The consequence is a lack of widely understood and accepted vocabulary to describe singing voice and vocal production which is at the same time a major challenge in studying it, thus reenforcing the vicious circle.

Singing has become a focus of attention in a variety of disciplines: singing as artistic expression, as an aspect of identity, gender and community building, as an acoustic phenomenon, vocalisation as a function of human body, vocal health and rehabilitation, psychological aspects of singing, etc. All these research fields

use different vocabularies. A Western singing teacher would often use subjective visualisations to describe vocal production: the sound can be round, warm, light, metallic, brassy, heavy, etc. Resonance would probably refer to a vague concept of louder and better projecting sound; while for acousticians resonance means partials amplified by the vocal tract. Voice scientists would mention formants, spectrum, amplitude, vocal folds closing rate, and other spectral descriptors, which can be studied on very short time frames (of several milliseconds – comparable to the length of a vocal folds vibration cycle) but are often hard to generalise for a longer time scale. Medical professionals – phoniatricians, surgeons, otolaryngologists and speech language therapists – use rating systems such as GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) (Little, Costello and Harries 2009) to assess voice disorders, or parameters like jitter, shimmer and noise-to-harmonics ratio. They also employ the vocabulary of vocal physiology – position of the larynx, thickness of the vocal folds, thyroid cartilage tilt, etc – to describe how the sound is produced in a singer's body.

Even within a single culture that has been thoroughly studied, such as Western music, there is little agreement among professionals about basic terminology for vocal production (Garnier et al. 2007a, Mitchell et al. 2003). Publications in English analysing vocal production in other cultures are rare (Födermayr 1971, Bartmann 1994). Singing teachers often use idiosyncratic language based on their subjective perception or learnt from their own teachers, therefore it is hard for teachers from different schools to agree about the terms (McGlashan 2013). Medical professionals are mainly interested in vocal dysfunction. Ethnomusicologists focus on the context of music making and rarely touch on vocal production itself; while for musicologists or music critics unique characteristics specific to the particular writer are considered an advantage.

Our interest in cross-cultural description of vocal production emerged from revisiting a seminal ethnomusicological study of singing performed by Alan Lomax and his Cantometrics team (Lomax 1968). Lomax conducted a large-scale venture to relate singing performance practices to societal traits by means of statistical methods: if, for example, in a given culture solo singing with a pressed and narrow sound, lots of embellishment and a heterophonic accompaniment was preferred, he concluded from his study that this society would have a stratified hierarchical structure and women would be subordinate to men (Table 1.1). The study was strongly criticised by ethnomusicologists – the methodology was in contradiction to the discipline's relativistic paradigm – and has never been systematically re-examined. At the same

time, in spite of its controversies, it remains popular outside ethnomusicology, and for some respected scholars in the field it is worth reconsidering (Merriam 1969, Nettl 2006). We wished to find out whether contemporary MIR approaches might open up new avenues to automate and scale up the Cantometrics experiment, thus overcoming some of the methodological weaknesses pointed out by the critics. It was the need to automate vocal production classification that directed our thinking towards objective vocabulary for vocal timbre.

1.1 Broad context and motivation – revising the Cantometrics experiment

Lomax approached the central question of ethnomusicology – the relationship between music (singing) and culture – by means of statistical methods. His idea was that singing is a mode of communication in society, a mode which is highly regulated – all society members know what is right or wrong, good or bad singing, and how one is to behave in a performance. Singing therefore is bound to reflect general communication patterns of the given society, which are in turn reflected in all other aspects of interaction and culture. In other words, if we know how people sing, we should be able to conclude about how they live together, how their society functions (Lomax 1977, p. 14ff). Lomax even went so far as to talk about a musical core of a culture – core cultural preferences for a music or singing performance – which he expected to be present in almost any musical utterance from the given culture (Lomax 1977, p. 16). In his Cantometrics experiment he was examining possible correlations between the musical descriptors and societal traits like stratification and child rearing (Lomax 1968).

The experiment was based on a dataset of over 5000 recordings from over 500 musical cultures. Performance practice was parametrised via 36 characteristics (see Table 1.1), some of which – *vocal width*, *rasp*, *nasality*, *volume*, *glottal shake*, etc – were directly related to vocal production.

Lomax's comparative, statistical methodology over a wide cross-cultural scope ran contrary to the paradigm of relativism and studying cultures from an emic[1] perspective that was prevalent in ethnomusicology at the time. The critique of Cantometrics by ethnomusicologists was overwhelming, the main methodological

[1] In anthropology, folkloristics, and the social and behavioral sciences, emic and etic refer to two kinds of field research done and viewpoints obtained: emic, from within the social group and etic, from outside.

weaknesses being the superficiality of the approach, and a small number of samples (usually ten) from each society in the dataset. It was stressed that many cultures have complex, multimodal musical traditions, which were not well represented in the Cantometrics dataset (O'Henry 1976); it was also pointed out that Cantometrics predictions of societal traits were not in accord with reality for a number of cultures (Feld 1984). Despite all the criticism there remain well-respected voices among ethnomusicologists saying that Lomax must have tapped into something really important (Nettl 2005).

Savage et al. (2015) conducted a contemporary study in the spirit of Cantometrics based on a revised set of musical/performance practice descriptors, their primary interest was to investigate musical universals. Yet it is telling that they omitted almost all descriptors of vocal production from the Cantometrics parametrisation system, with the exception of pitch/interval size and loudness (the majority of performance practice descriptors were omitted as well).

No attempts have been made to our knowledge to replicate the anthropological component of the Cantometrics experiment on other data; or to revise it addressing its methodological weaknesses such as subjectivity of terms, raters diversification, data and scalability limitations. Some of these weaknesses might be attributed to the lack of technological infrastructure at the time of the experiment; others require careful scientific consideration.

The data from the Cantometrics experiment was never published in its entirety. The scope of the experiment was huge for the 1960s. Audio recordings were physically copied on audio cassettes (often from more obscure media such as reel tapes); digital repositories as we know them today did not exist. Lomax was determined to publish the recordings in the form of what he called "A Global Jukebox" – a playlist of songs from every culture in the world. He imagined it to be an educational resource as well as entertainment (at the time when no world music radio stations yet existed), giving every person access to the music of their ancestors, to be proud of their heritage and appreciative of other cultures. Moreover, because all the songs were annotated with 36 parameters of performance practice, it would in fact constitute a database allowing for targeted searches and comparisons. A tireless advocate of "unrecorded and unrecognised music", Lomax tried to influence US government policy on arts and culture on every level. His work fed into the influential UNESCO "Appeal for Cultural Equity" in 1972 arguing the right of every culture to safeguard, express, and develop its artistic and expressive heritage. In 1986 Lomax founded the Association for Cultural Equity "to explore and preserve the world's expressive

traditions with humanistic commitment and scientific engagement"[2].[3]

What we call the Cantometrics dataset – audio recordings annotated with the Cantometrics parameters and used for statistical investigation of relationships to societal traits – does not exist as an entity today. Lomax's own field recordings were included as well as material from hundreds of other collections. The enormously complicated legal rights handling makes its reassembling a mammoth task.

In 2006 Anna Lomax Wood kindly provided us with the then recently digitised collection – the Cantometrics Training Tapes. The recordings are a subset of the Cantometrics dataset of several hundred tracks. They were used as illustrations and test examples for the textbook on Cantometrics Lomax and his colleagues developed, which was also used to train raters (Lomax 1977). This collection has been the subject of our research ever since: testing and developing general MIR techniques, analysing why and where they failed for such a varied dataset (Proutskova and Casey 2009); collaborating on a development of a novel dynamic user interface to access this collection, including similarity search, loosely following on Lomax's Global Jukebox idea (Magas and Proutskova 2013). Musical examples in this PhD are taken from the Training Tapes (see Section 4.3).

Victor Grauer, an established ethnomusicologist and Lomax's main collaborator on the Cantometrics project suggested to concentrate on one Cantometrics descriptor – *vocal width/vocal tension,* which in the original experiment was found to correlate to subordination of women (Lomax 1977, pp. 26-27 and pp. 125-126). Lomax discovered that in societies where narrow, tense, squeezed vocal production was the norm, the pre-marital sex sanctions for women were more severe than in societies where wide, open, resonant vocalisation was preferred. This statistical result supported Lomax's insight from his field trips to Spain – while the Andalusian South was under a strong Islamic influence, where women were hidden behind burkas and the preferred vocal sound could be described as tense and narrow, in the far more relaxed North people also sang very differently, in big, resonant, relaxed voices. This insight was his starting point for the whole Cantometrics endeavour (Lomax 1968) and one of the most controversial Cantometrics findings.

[2] `http://www.culturalequity.org/ace/ce_ace_index.php`, last accessed on 10 September 2017

[3] Unfortunately, the lack in the 1960-70s of the digital infrastructure that is ubiquitous today made the realisation of his project very difficult – the Cantometrics endeavour was too much ahead of its time. By the time when the infrastructure became available 40-50 years later, several generations of data formats had come and gone and public interest in the project had waned. Lomax's original recordings are preserved at the Library of Congress. The Association for Cultural Equity led by his daughter Anna Lomax Wood continues to digitise and gradually publish Lomax's extensive heritage.

Table 1.1: Correlations between musical and societal parameters discovered in Cantometrics. This table gives an overview of general relationships for groups of Cantometrics parameters (factors). For more details on correlations see Lomax 1977, pp. 22-28 and 260-269.

Musical factor	Parameters included*	Societal descriptor	# cultures	p-value
Differentiation (information load)	Enunciation, repetition of text, interval size	Productive scale	157	.001
Ornamentation	Melisma, glissando, embellishment, glottal	Large domesticated animals	97	.001
Orchestral organisation	Musical and social organisation of the orchestra	States	82	.001
Cohesiveness	Vocal blend, rhythmic coordination	Community solidarity	143	.001
Choral organisation	Solo/group, Musical and social organisation of the vocal group	Solidarity	102	.01
Noise/Tension in voice production	Nasality, rasp, vocal width	Severity of sex sanctions	117	.001
Energy level/ dynamics	volume, accent, pitch	Extra-local government hierarchies	151	.001
Irregular to regular rhythm	Rhythm (vocal, orchestral), tempo, melodic variation	Infant/child indulgence	40	.001/.01
Melody (complex/simple)	Melody form, number of phrases, phrase symmetry	Large/small settlement	124	.001

Following Victor Grauer's fascinating investigation into the evolution of human singing and music-making style from its very beginning (Grauer 2006a)[4], we felt that this line of enquiry would provide an excellent collaboration opportunity for modern technology (e.g. MIR) and voice science, with results having relevance in a wide variety of fields. The striking, speculative, counter-intuitive but tempting idea of women's status in the society being related to vocal timbre preferences served us as the underlying motivation to approach vocal production with contemporary techniques and to re-visit Cantometrics. Thanks to Victor Grauer's supportive encouragement we started looking into ways to use MIR to automatically extract Cantometrics parameters from audio recordings.

What if we could analyse all recorded singing instead of just several thousand samples in the original Cantometrics dataset? That would address the small number of samples per culture issue and might lead to new and unexpected insights. This approach would require automatic analysis of audio recordings instead of manual ratings: in the original Cantometrics experiment each sample was analysed by at least three raters who provided values for the musical descriptors; while this task was manageable for 5000 recordings, it would be impossible to realise for an open dataset with new recordings added on regular basis. A proper training for automatic annotation model would offer an opportunity to reduce the cultural bias present in the original ratings – all Lomax's raters were American ethnomusicology students.

Machine learning (bottom-up) approaches to audio analysis such as automatic extraction of high-level descriptors (like Cantometrics *vocal tension*) require a significant corpus of recordings annotated with the descriptor in question. This brings us back to the subject of vocabulary: there are no commonly understood and widely used terms for vocal production, therefore there are no corpora annotated for vocal production. Musicological (top-down) computational approaches could be based on a formalised knowledge of vocal production, in particular of vocal tension, but formalising knowledge would again require formalised or at least widely accepted vocabulary. For his experiment Lomax had to invent his own descriptive system, which he formalised to the extent, necessary for his study, but not enough for computational modelling: the raters were trained on the basis of his textual descriptions

[4] In a truly vast attempt to outline his views of the global history of human musical style, its origins and evolution, Victor Grauer (2006a) (the co-inventor of Cantometrics) relies heavily on Cantometrics analysis and on his experience of working with Alan Lomax on the Cantometrics project. He also draws on modern genetic, archaeological and linguistic research (Grauer 2007). Publication of this work in the World Of Music journal caused a lively discussion and resulted in two issues of the journal devoted exclusively to this subject (Nettl 2006; Stock 2006; Cooke 2006; Grauer 2006b; Rahaim 2006; Cross 2006; Mundy 2006).

and musical examples. He was not a singer and his definitions of vocal production, at least in case of *vocal tension*, were simplistic and in some cases led to contradictions (Section 2.3).

Our hypothesis at the start of this PhD was that we can find more objective, more formal descriptors of vocal production and re-model *vocal width/vocal tension* by mapping it onto our descriptor space. These objective descriptors would allow us to build computational models to automate their identification in audio recordings. For objective descriptors diversification of raters will not be necessary; and the number of samples per culture could easily be increased. The statistical correlation between the mapping of *vocal tension* on our descriptor space on the one hand and subordination of women (sanctions for pre-marital sex for women) on the other could then be re-investigated on a larger dataset than the Cantometrics one.

1.2 Vocal production vocabulary

There is no established terminology describing vocal production, neither a widely accepted vocabulary, not even within a single, well studied musical tradition.

Vocal health seems to be the field with the most ordered approach to describing vocal production. While objective parameters such as jitter, shimmer and harmonic-to-noise ratio can be measured with the hardware that is available at many clinics, auditory-perceptual scales (GRBAS, CAPE-V) are still most popular (Oates 2009). They often employ terms borrowed from everyday life (such as strain, roughness, breathiness) with which clients feel more comfortable than technical characteristics. Yet acceptance and familiarity do not guarantee objectivity, instead they often mask polysemy and various connotations; experienced voice clinicians rarely agree about the values of perceptual terms (Kreiman et al. 1993).

In singing education the situation is the exact opposite: there are no established perceptual scales, neither are any direct measurements generally available. Singing teachers more often than not use idiosyncratic vocabulary which either is borrowed from their teachers or illustrates their own imagery. Physiological precision is very rarely a goal. Even most widely used terms vary in their meaning considerably – see e.g. Mitchell's et al. (2003) semantic analysis of the term open throat.

There are singers' and teachers' communities that employ some agreed terminology. Garnier (2007b) analyses the vocabulary of classical singing teachers in France: of around 600 terms most are unique to the teacher, often metaphorical and onomatopoetic forms. There is a set of common terms of about 30 which were not

unique and surfaced in many interviews, sometimes they were explicitly referred to as common knowledge in the profession. Several of these standard terms were polysemic, displaying a number of different meanings with sometimes opposite connotations (Garnier et al. 2007b). Another example of a community terminology is Catherine Sadolin's *Complete Vocal Technique* (2000) which offers its very unique set of terms (such as *curbing*, *edge*, *hold*); these are used by her community of international followers. Yet understanding beyond community boundaries or between various schools of teaching is hindered by ambiguous and imprecise terminology (McGlashan 2013).

We present here two examples of terminology used by singing teachers to describe vocal production. The first list is a collection of various descriptions of a belt sound which is extensively used in music theatre and contemporary commercial music (from Jeannie LoVetri's/Somatic Voicework Teachers Association blog[5]):

Belt: light, brassy, twangy, forward, heavy, warm, chesty, lyrical, ringy, bright, thick, shouty, whingy, mix, etc.

All these descriptors are subjective, meaning different things for different people. The other example is a list of names for "voice gears", vocal mechanisms or registers used by various teachers and communities in Western singing education (from VocalProcess teaching webinars[6])

- Chest, Head, Mix, Falsetto
- Mechanism 1 & 2
- Modal Voice & Loft
- Heavy Mechanism, Light Mechanism
- Thick, Thin folds, 'Stiff' & 'Slack'
- Shortener (TA)-dominant & Lengthener (CT)-dominant
- Neutral, Curbing, Overdrive, Edge

These terms refer in principle to the main characteristic of vocal production – the vocal folds vibration. Yet there is no common word for it that would be accepted by

[5] `http://somaticvoicework.com/category/jeanie-lovetri-blog/`, last accessed 31 August 2017.

[6] `http://store.vocalprocess.co.uk/Webinars`, last accessed 31 August 2017.

most communities; there is no clarity to what extent the terms in the list describe the same things, where exactly the differences lie and how the descriptors correspond to physical reality.

While singing teachers' main aim is to help their students sing better, Alan Lomax's goal when he devised Cantometrics parameters was classification of singing performance practice. He later grouped his parameters to statistical factors which miraculously represented common meaning and correlated to the same societal traits (Lomax 1977). The factors related to vocal production were:

- Noise/tension factor: rasp, vocal width/vocal tension, nasality
- Energy level factor: volume, accent
- Ornamentation factor: glottal shake, tremolo, glissando, melisma, embellishment

Lomax believed his parameters to be universal in the sense that anyone could be taught to understand and rate them through a short training. Yet the evidence is insufficient due to lacking diversification of raters.

Our aim is to develop a more formal language about vocal production that would allow us to map different reflections of the process of vocal production onto each other. We would like to employ the analogy to the physical maps: all of us who work with the voice have our maps of the vocal production process. These maps can vary in the amount of detail, in the emphasis, they can represent different aspects of the voice. Yet we know less about the process of singing than about the Earth's surface. The maps of the Earth can all be checked against satellite pictures; if a map does not match them correctly, the map is wrong. There are no satellite pictures for singers' bodies, therefore the validity of our vocal maps cannot be verified. If two physical maps represent the same bit of the Earth, they can be mapped onto each other. But would our vocal maps, reflecting the same aspect of vocal production, agree?

1.3 Models of vocal production and our ontology

There is no theoretical model of vocal production which could provide the basis for predictions. There are no annotated datasets either. As we have seen above (Section 1.2), there isn't even a vocabulary to talk about vocal production. We evaluated three approaches to parametrising vocal production that have had a wider

reach: the Cantometrics study, originating in anthropology and ethnomusicology; Jo Estill's physiological "building blocks" system that has been influential in singing education; and voice source characteristics such as phonation modes as described by Johan Sundberg, the founder of singing voice science.

While the Cantometrics approach was the one we wanted to re-examine, in particular due to subjectivity of the language that was used, we turned to vocal physiology and voice science for a more objective description of vocal production.

While terms *breathy, neutral* and *pressed* voice had been used by speech voice researcher as well as voice therapists to describe phonation Laver 1987, it was Johan Sundberg, the founder of the singing voice science, who formalised them for the singing voice in his seminal volume "The science of singing voice" (Sundberg 1987). He related them to the aerodynamic processes such as vocal resistance as well as to the vibration patterns of the vocal folds.

Jo Estill was an American singer, teacher and voice researcher, who suggested a physiology-based system for understanding and teaching vocal production. Her idea was to isolate physiological structures, learn to control them independently and use these building blocks of vocal physiology to construct various kinds of vocal production, ultimately leading to the ability to build any singing style (Estill and Colton 1979, Colton and Estill 1981). Her work had a huge impact on contemporary singing education (Sadolin 2000, Soto-Morettini 2006, Kayes 2004).

Since we could not verify the inter-personal and inter-cultural consistency of the Cantometrics approach we concentrated on the physics and the physiology of the sound. Our approach is based on the analysis of vocal source and vocal tract settings. Vocal source setting is the laryngeal mechanism of sound production which includes the aerodynamic process of vocal folds oscillation and air wave propagation as well as the physiological configuration of the larynx; vocal tract setting is defined by Laver as a "long-term average configuration of the vocal organs ... underlying momentary segmental articulations" (Laver 1980, p. 10). The study of Sundberg's vocal source research (Section 2.1.1), the classical Western registration (Section 2.1.2) and the Estill model (2.1.3) ultimately led to a compilation of a vocal production ontology (Section 2.2).

1.4 Revising Cantometrics: methodological challenges for MIR

The relationship between objective (physiological) and subjective (perceptual) description of vocal production refers to one of MIR's challenges – the semantic gap (Celma and Serra 2008, see also Wiggins 2009). We suggest to follow the path of introducing a middle layer of objective, measurable descriptors in an attempt to model a high level characteristic (Cantometrics parameters) via low-level information from the audio signal.

Embarking on the journey of revising the Cantometrics experiment using contemporary technological advances the main challenge we face is the lack of annotated datasets of vocal production – these are needed to train and test computational models. The reasons for the lack of such a dataset lie outside MIR and have been discussed in the previous sections: insufficient knowledge about the mechanisms of vocal production; objective measurement very limited and restricted to real-time contexts; no widely accepted vocabulary. Because of the fundamental nature of the above challenge this thesis is dedicated to investigating possible approaches to overcome it.

Cantometrics data

The Cantometrics dataset contains recordings from all around the world. There is a huge variation in musical content – in fact the dataset was compiled to represent all the cultural variation in musical style present in our human culture in the middle of the 20th Century (Proutskova and Casey 2009, Magas and Proutskova 2013). Cantometrics measures this variation along 36 musical style descriptors (see Section 1.1). In MIR related terms, there are monophonic as well as polyphonic recordings, solo and group singing, male, female, children's and mixed group singing. Singing can be a cappella as well as accompanied, and the orchestras accompanying singers include all kinds of instruments. Various rhythms and metres are present including polyrhythms and non-metric pieces. There are recordings in scales that differ from the Western tempered scale.

There is also a considerable variation in recording conditions. Many recordings in the Cantometrics dataset were made in field conditions and originate from the first half of the 20th Century. They can be very noisy. They can also contain sounds from the environment, such as nature sounds or musicians speaking during

performance. Others are studio recordings from a time period spanning 40 years and more

All sorts of audio formats and compression will have to be dealt with. While modern recordings can be as good as 128 kHz and 32 bit precision uncompressed, older recordings are most certainly of a lower resolution. Digitisation of analogue recordings was performed by various parties to differing specifications.

We are therefore faced with one of the most general MIR tasks: automatic extraction of a high-level descriptor from a highly heterogeneous audio dataset.

Incremental vs integrated approach

Two different approaches to this complex task are considered in our thesis. The first is an incremental approach, when we begin with a manageable MIR problem based on controlled data and in each following step this problem is generalised to a set of data containing more variation. The integrated approach is based on collecting reliable annotations for the original, highly varied data.

In particular, we begin with introducing a simplification. The question of automatic vocal production recognition is reduced to one descriptor: phonation mode. Instead of real-life ethnomusicological recordings we look at sustained vowels recorded under controlled lab conditions. We create our own dataset with annotations for the task and devise a computational model to recognise phonation modes automatically. This work is presented in Chapter 3. The generalisation and its challenges are discussed in Section 3.4 of that Chapter.

The second approach we investigate in this thesis avoids simplification and addresses the variability in the data head on. It aims to produce reliable annotations of vocal production for an existing dataset – the Cantometrics recordings – which comprises all the variability aspects listed above. In absence of direct measurements we turn to experts' knowledge elicitation. We conduct an investigative mixed-method study in which experts analyse recordings from the Cantometrics dataset to a) verify our ontology, b) scrutinise the experts' consensus and c) collect annotations (Chapter 4).

1.5 Document structure

Chapter 2 begins with a review of existing models of vocal production and vocal physiology. First, Sundberg's research on voice source and phonation modes is

presented (Section 2.1.1); then the development of the classical registration theory is outlined (Section 2.1.2); following that we touch upon Jo Estill's impact on singing education and research (Section 2.1.3.1) and explain the Estill model of physiological building blocks (Section 2.1.3.2).

The chapter continues with a compilation of a vocal production ontology that includes previously discussed descriptors (Section 2.2); the choice of scales is justified. We then return to the Cantometrics descriptors, in particular to *vocal width*. This descriptor is deconstructed using the knowledge gained from the background review, demonstrating that its components are not directly related as Lomax assumed (Section 2.3). Then we hypothesise about which ontological descriptors could contribute to the perception of the vocal width's components (Section 2.4).

Chapter 3 is dedicated to the incremental approach to automatic vocal production analysis. It presents our experiment on automatic classification of phonation modes in recordings of sustained vowels. Details are given on the low-level feature extraction based on glottal waveform reconstruction via inverse filtering (3.1.1). The Phonation Modes Dataset we produced for this experiment is described (3.2), as well as the experiment design (3.1) and results (Section 3.3). Then we discuss generalisation steps that would be required to replicate the experiment on more varied data (Section 3.4). Challenges related to the generalisation process are explained.

Chapter 4 follows the integrated approach to automatic vocal production. It lays out a methodology for creating reliable annotations for the Cantometrics dataset. A mixed-method study of expert knowledge elicitation is presented on a subset of 11 tracks (Section 4.1). 13 experts in various fields (medical professionals, singing teachers, voice scientists) analysed physiologically stable fragments of the 11 tracks in semi-structured interviews. Negotiating the interview time, number of tracks (Section 4.2), ontological descriptors (Section 4.4) is chronicled as well as musical examples preparation (Section 4.3). Data collection is detailed (Section 4.6) and strategies for two possible results scenarios are considered (Section 4.7). We then document the iterative process of the experiment progression (Section 4.8).

Chapter 5 deals with the quantitative analysis of collected data. It addresses the acceptance of the ontology descriptors and the study design by the participants (Section 5.1). Then inter-rater agreement measure called *Krippendorff's alpha* (Section 5.3) is calculated for each physiological descriptor (Section 5.5) and its confidence interval is computed by means of our own R routine (Section 5.4). The limitations of Krippendorff's measure in relation to sparse data are investigated (Section 5.6) and our extension of the Krippendorff's algorithm to weighted ratings is described

(Section 5.7). For the ontological descriptors which displayed inter-rater agreement reliable annotations are collected (Section 5.8) to be published with the first cross-cultural dataset of vocal production. Also, statistical correlations are established between these descriptors and the Cantometrics parameter *vocal width*. A follow-up experiment is proposed that would exclude cultural variation (Section 5.10).

In the following Chapter 6 the qualitative analysis of more than 33 hours of interview recordings is detailed. It scrutinises participants' views on each of the rated descriptors. Then their suggestions for additions or changes to the ontology are considered (Section 6.9).

Based on insights and conclusions from qualitative analysis (see Section 6.10) meta-analysis of confounding issues and common themes is performed in Chapter 7. It identifies six themes which point to the reasons for disagreement between our expert participants. In Chapter 8 these themes are discussed from the viewpoint of MIR (Section 8.2) and vocal pedagogy (Section 8.3) with conclusions and further research suggestions in each of the fields. Our progress towards revising and scaling up Cantometrics is discussed (Section 8.1) including the success of mapping *vocal width* onto more objective descriptors (Section 8.1.1) and the failure of general physiological approach (8.1.4). A mixed objective-subjective ontological concept for revising Cantometrics is explored (Sections 8.1.3 and 8.1.6) and an updated research plan for revising and scaling up Cantometrics is suggested (Section 8.1.6.5).

2 Ontology

In Section 1.2 of the Introduction we discussed how various communities of voice experts use vocabulary on vocal production, how unsystematic and subjective these vocabularies are and how understanding beyond community boundaries is hindered. We aim to compile a more formal and more objective set of terms that would allow for building computational models and automatic classification applications.

In Section 1.3 three systematic approaches to vocal production were outlined – a perceptual one, from the perspective of performance practice and cultural preferences (Cantometrics); one based on voice aerodynamics and acoustics (Sundberg), and one proposing physiological building blocks to construct various sounds and singing styles (Estill).

This chapter provides a detailed study of Johan Sundberg's work on voice source (Section 2.1.1), the classical Western registration (Section 2.1.2) and the Estill model (Section 2.1.3). Based on this knowledge and terminology, an ontology of vocal production is complied, which comprises vocal source, laryngeal and vocal tract physiological configurations (Section 2.2). We then use the acquired knowledge and terminology to examine the Cantometrics descriptor *vocal width*, to untangle its components/dimensions and to uncover its underlying contradictions (Section 2.3). We also hypothesise about possible contributions by the aerodynamic and physiological factors to the perception of *vocal width*'s dimensions (Section 2.4).

2.1 Related previous work

2.1.1 Vocal source – Johan Sundberg

The voice organ consists of the lungs, the larynx, the pharynx and the mouth (Figure 2.1.1). The lungs generate an airstream that passes through the *glottis* – the area at the bottom of the larynx between the vocal folds – and sets them in vibration. Oscillating vocal folds in turn set the airstream above them in vibration: through frequent opening and closing they chunk the airstream into pulses. The vibrating airstream is called *vocal source*; because the vibration is periodic, the multiples of

Figure 2.1.1: Voice organ. Image by The Voice Clinic of Indiana, `http://www.voiceindy.com/anatomy-physiology-background/`

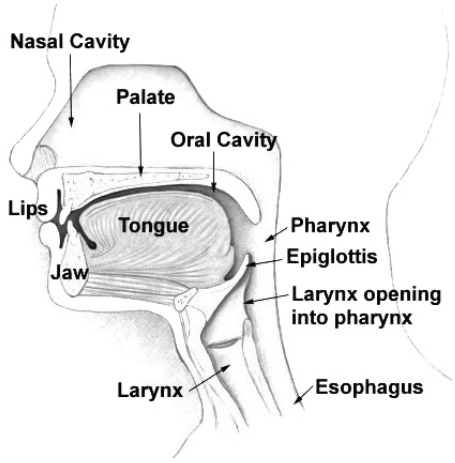


Figure 2.1.2: Head and neck structures. Image by Arcadian, `http://training.seer.cancer.gov/head-neck/anatomy/overview.html`, Public Domain.

the fundamental frequency are also produced, they are called *harmonics* or *partials*. The vibrating airstream travels through the pharynx, the mouth and nose cavities (Figure 2.1.2), where it is altered (voice scientists say *filtered*) by the form of the vocal tract with some partials enhanced and others weakened; in physical terms, vocal folds act as an oscillator and the pharynx and the mouth as a resonator (Figure 2.1.3). The vibrating air is then radiated from the lips and propagated to the ears of the listener. Vocal source largely defines the fundamental frequency (*F0* – pitch) and the sound pressure level (*SPL* – loudness) of the sound; vocal tract mainly determines its specific colour or timbre (though it can also contribute to the loudness).

Subglottal pressure and transglottal airflow are the main characteristics of vocal source aerodynamics: *subglottal pressure* is the pressure built under the vocal folds during the closed phase of the vibration cycle; *transglottal airflow* is the amount of air escaping the vocal folds during the open phase.

Johan Sundberg worked extensively on the aerodynamics of voice source (Sun-

Figure 2.1.3: Sound production (Sundberg 1977 p. 107)

dberg 1987, chapter 4). In that chapter he demonstrated that subglottal pressure has a strong effect on SPL or loudness of vocal production. He also introduced the notion of four phonation modes in singing: *breathy*, *neutral*, *pressed* and *flow*. These phonation modes result from different configurations of the voice source according to Sundberg. He related them to various vocal folds closure and opening patterns (Figure 2.1.4). Breathy and pressed phonation types are also used widely in other fields of research such as linguistics and vocal health (Laver 1987).

2.1.1.1 Phonation modes in singing: voice acoustics

The term *phonation mode*s was coined by Johan Sundberg. In his classic book "The Science Of The Singing Voice" (1987) he introduced four phonation modes: *breathy*, *neutral*, *flow* and *pressed*. They are vocal production qualities resulting from the voice source (the vibrating vocal folds). In particular they are closely related to glottal resistance which is defined as the quotient of subglottal pressure to glottal airflow. Generally speaking the phonation modes correspond loosely to regions in the 2-D space spanned by glottal airflow and subglottal pressure: a low subglottal pressure combined with a high glottal flow results in a breathy phonation; pressed phonation arises when a high subglottal pressure is accompanied by a low glottal flow.

In reality not all points of the above 2-D space can be realised physically. Each singer is capable of vocal production in a subspace depending on the nature of their voice apparatus, their habits and their training.

Phonation modes can be illustrated by means of the typical voice source signal waveforms. The graphs in Figure 2.1.4 are taken from Sundberg's book (1987, p. 85); they show one full cycle of vocal fold vibration: beginning with the closed phase, when no or little air escapes the vocal folds, followed by the opening phase when the vocal folds part and let through a stream of air. Pressed phonation displays a long closed phase, with reduced airflow during the opening phase. In the neutral mode the closed phase is somewhat shortened and the airflow during the opening phase is considerably increased. This trend is continued in the flow phonation, with a still shorter, though evident, closed phase followed by an opening phase with high glottal airflow. In the breathy vocalisation the airflow is raised further, and the closed phase virtually disappears: the vocal folds never close completely, which leads to the leakage of air at any time during the cycle. The subglottal pressure is high for the pressed sound, approximately average for the neutral and the flow

sounds and low for the breathy.

Sundberg introduced flow phonation in search of a descriptor that would help explain the prevalence and particular qualities of Western classical singing. He described flow phonation as the sweet spot where the maximal airflow is achieved retaining a closure of the vocal folds during the closed phase. In particular the flow phonation usually displays a lower subglottal pressure than the pressed mode and also a lower airflow than a breathy sound. This makes the flow phonation an economical voice production mode, requiring less physical effort (less pressure, less air) than both pressed and breathy modes. The flow model seems to reflect the practice of classical singing which requires an extensive breath supply and control, does not employ the levels of effort of pressed voice and allows for volume levels which are unachievable in breathy phonation. Another characteristic of flow is a high fundamental and a slowly declining spectrum slope, as opposed to a quick decline in breathy and neutral or a weaker fundamental in pressed.

Phonation modes defined by Sundberg thus describe the distinctive vocal fold closure and opening patterns. This term does not refer to the differences in phonation between the modal and the falsetto registers (M1 and M2, see section 2.1.2), when different vibratory mechanisms of the vocal folds are involved. Sundberg does not mention registration; in fact, the question remains open whether all phonation modes can be realised in both M1 and M2 for male and female singers, and in which ranges. While breathy or neutral sound can usually be produced in all ranges even by less experienced singers, pressed may not be accessible in M2 (possibly because too high subglottal pressure would interfere with vocal folds oscillation); employing flow usually requires training, particularly in order to widen its range. While much of Western operatic singing for female singers takes place in M2, particularly for sopranos (Gillyanne Kayes, personal communication, October 2017), this is not the case for male singers. This difference may be an indication of the varying ranges/mechanisms in which flow phonation can be achieved for males and female singers respectively.

Several studies have been published attempting to determine dominant phonation modes or typical values of glottal flow waveform descriptors for various singing styles. For example Thalén and Sundberg (2001) and Sundberg et. al. (2004) studied Western classical music, pop, jazz and blues. A female singer sung a triad pattern in four phonation modes as well as in the above singing styles. Various glottal flow waveform derived measures of glottal adduction were analysed in their relationship to perceived phonatory pressedness, including Normalised Amplitude

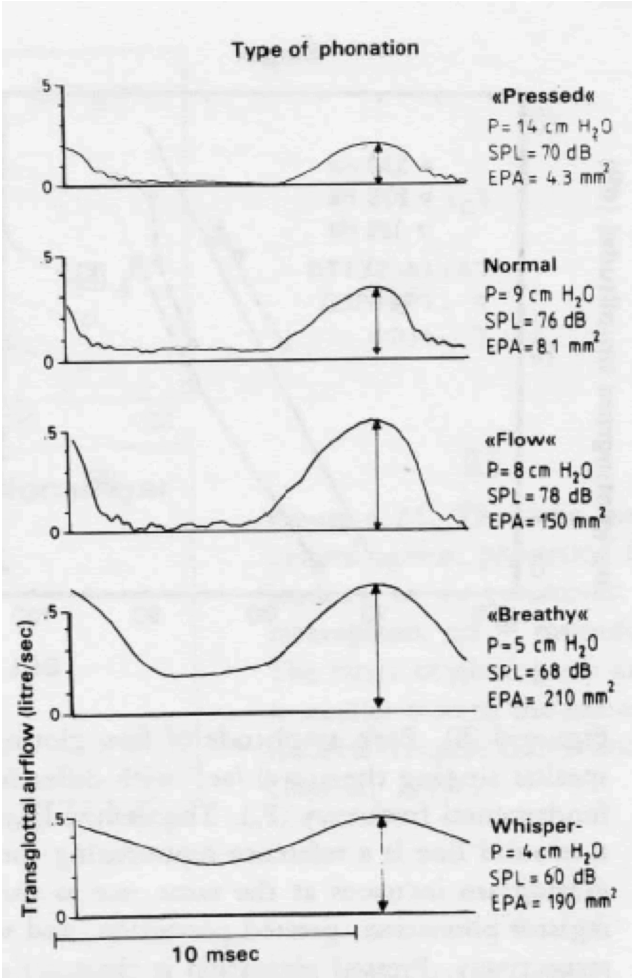


Figure 2.1.4: Typical graphs of the glottal flow waveform pulse functions in various phonation modes (from Sundberg 1987, p. 85)

One full cycle of the vocal folds vibration is shown: beginning with the closed phase, when no or little air escapes the vocal folds, followed by the opening phase when the vocal folds part and let through a stream of air. On the right side the values for subglottal pressure P (measured by means of the Rothenberg mask), the signal pressure level SPL as well as for the transglottal airflow amplitude maximum EPA are given.

Printed with permittion from Prof. Sundberg.

Quotient (NAQ), the difference between the first and the second harmonics (H1-H2) and the closed quotient (ClQ). NAQ was found to account for over 70% of variation in perceived pressedness. Also, samples of blues singing were perceived by a panel of experts to be the most pressed, in contrast to classical singing with least pressedness, pop and jazz residing in between. Mean subglottal pressure for blues samples was higher than for other styles. The values of mean NAQ were found to differentiate well between the styles of the samples.

In a later publication Borch and Sundberg (2011) looked at rock, pop, soul and Swedish dance. Here the setting was closer to real life recordings: beyond the triad patterns, a male singer sung songs in the named styles. In contrast to the previous work, it was found that the mean NAQ values were similar among these singing styles. This might be accounted for by the differences in range and loudness between those styles: e.g., rock singing was expected to correspond to a lower NAQ due to more pressedness, but at the same time it was sung on higher pitches, which in turn correspond to higher NAQ values. Regarding subglottal pressure, rock displayed the highest values in contrast to low pressure in Swedish dance, with pop and soul again residing in the middle. Also, significant differences between styles were found in the long-term average spectrum (LTAS).

These studies worked with recordings by just one singer. As a starting point this approach is instructive. Unfortunately, the methodology suggested in these papers does not scale to large datasets and batch processing applications. Also, the data on which the results are based was not made available to other researchers, thus making direct comparisons as well as iterative dataset expansion and methodology improvements by others impossible.

2.1.1.2 Performance practice

In this Section we provide examples of uses of phonation modes from various sources to illustrate the differences between them.

Breathy vocalisation is used skilfully by jazz and popular music singers to express qualities like sweetness or sexuality: think of Marilyn Monroe's most famous performances like "I wanna be loved by you"[1] or "Happy birthday Mr President"[2]; or listen to Chet Baker's singing, such as "My funny Valentine"[3]. This mode of vocal production can easily be distinguished by human listeners from the flow phonation

[1] http://www.youtube.com/watch?v=MLU0jndUGg4 (last accessed on 30/10/2012)
[2] http://www.youtube.com/watch?v=k4SLSlSmW74 (last accessed on 30/10/2012)
[3] http://www.youtube.com/watch?v=7iQQGBfbB0k (last accessed on 30/10/2012)

mode, such an operatic baritone or soprano voice; or from the pressed phonation, e.g., the tense, forceful voice of James Brown in "I feel good"[4].

While the term phonation mode is borrowed from voice acoustics, the differentiation between breathy and pressed voices, between tense and open singing is operational in many voice-related research areas: singing education, medical research (phoniatrics, vocology), linguistics (phonetics) as well as in singing performance. The use of breathy, pressed or resonant singing production can be representative of an individual singing style as well as of a particular musical repertoire. While each voice is different and two singers never sing the same way, repertoires within a music tradition (or sometimes across music traditions) display cultural preferences for the use of particular phonation mode(s), which are imposed on the singers performing in these repertoires. For example baritone singers in Western operatic repertoire are trained to sing in flow phonation (using the neutral mode occasionally to cover the register break) and move through their singing career using just this phonation mode. In contrast, in the classical Ottoman tradition a singer is expected to operate in all four phonation modes.

Apart from being a stylistic characteristic, breathy or tense vocalisation can be indicative of vocal disorders: hypofunction and hyperfunction of the glottis (Froeschels 1943). Their diagnostics and treatment are a prime concern in voice rehabilitation and phoniatrics (in case of functional or anatomic pathologies) (Ramig and Verdolini 1998).

Voice therapists specialise in vocal production and could therefore serve as expert listeners for manual rating of phonation modes. In practice, though, their work is often tailored more to the needs of speech professionals. In singing it is singing teachers/educators who have the deepest operational knowledge of all the issues related to vocal production and in particular to phonation modes. Most singing students display various kinds of voice hypo- and/or hyperfunction during the stages of their progress (Froeschels 1943). The students' perception mechanisms are not sufficient for self-control (in absence of any visual or any reliable auditory indicators). It is therefore the task of the teacher to identify and to correct the subtlest dysfunction, over and over again, until the student has gained the bodily controls necessary to regulate the voice source function on an automatic level.

[4]http://www.youtube.com/watch?v=XgDrJ5Z2rKw (last accessed on 30/10/2012)

2.1.2 Registers

Vocal registers as a perceptual phenomenon of different voice qualities depending on the range have been known in the Western classical vocal tradition for a long time. It was the pioneering singing teacher Manuel Garcia Jr. (1805-1906), the first to successfully observe the larynx during singing by means of a mirror laryngoscope (García 1855), who laid out physiological differences as the basis for distinguishing registers. Garcia's famous definition of registers has coined the understanding of the phenomenon as well as the confusion related to the term until today. In his presentation to the French Académie des Sciences on 16 November 1840 he stated:

> "By the word register we mean a series of consecutive and homogeneous tones going from low to high, produced by the same mechanical principle, and whose nature differs essentially from another series of tones equally consecutive and homogeneous produced by another mechanical principle. All the tones belonging to the same register are consequently of the same nature, whatever may be the modifications of timbre or of the force to which one subjects them." (Garcia 1847)

In the paper "Ecole de Garcia: Traitécomplet de l'art du chant" that lays out his singing voice teaching method (1884) he refers to three registers: *poitrine* (chest), *fausset-tête* (falsetto-head), and *contre-basse* (counter bass). Through his observation of larynx position he found that head and falsetto (he placed the latter between head and chest) were based on the same laryngeal mechanism, differing only in timbre.

As Roubeau et al (2009a) note, this definition, which remains the most widely cited description of registration, however coherent to Garcia himself, has been a source of ambiguity for generations of researchers. The notion of homogeneity he evokes refers to mechanical principle in his definition. Firstly, the term homogeneous has been interpreted in numerous ways by different authors, often as a reference to perceptual similarity. Secondly, he does not define the mechanical principle that underlies different registers thus leaving more opportunity for confusion (Roubeau, Henrich and Castellengo 2009b).

In 1880 the mechanical principle was spelled out by the physiologist and singing teacher Emil Behnke and the throat surgeon Lennox Browne who also made use of the laryngoscope to obtain in vivo images of the glottis (Behnke 1880). Behnke defines the vocal registers as follows:

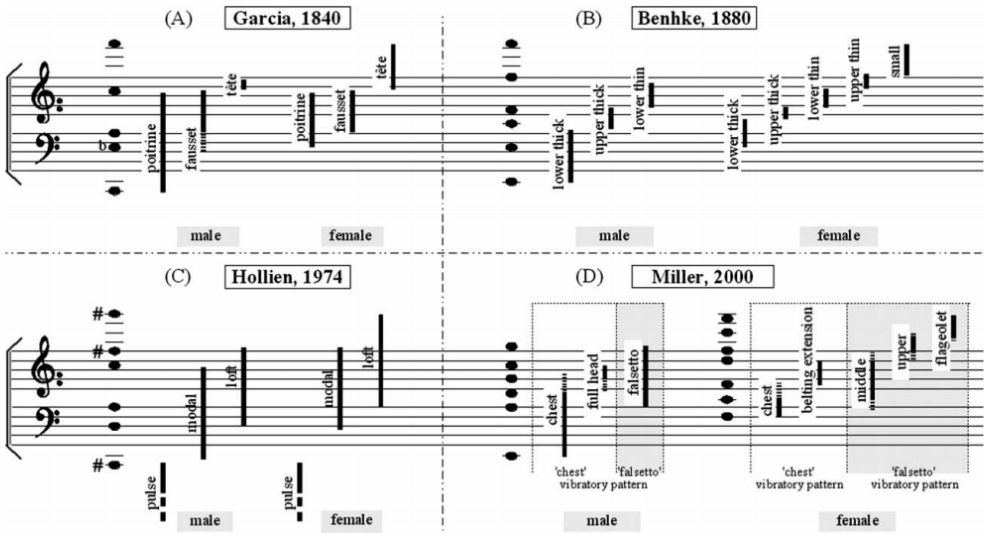


Figure 2.1.5: Frequency range of human voice and vocal registers, as defined by different authors: (A) Garcia (1), (B) Behnke (4), (C) Hollien (11), (D) Miller (17). From Henrich 2006 with permission from Taylor & Francis.

> "... a register consists of a series of tones which are produced by the same mechanism. [...] There are, broadly speaking, three registers in the human voice, and the mechanisms are plainly visible, as follows: 1) During the lowest series of tones the vocal ligaments vibrate in their entire thickness. 2) During the next series of tones the vocal ligaments vibrate only with their thin inner edges. 3) During the highest series of tones a portion of the vocal chink is firmly closed, and only a small part of the vocal ligaments vibrates." (Behnke 1880, cited after Henrich 2006)

On the basis of these physiological observations, he adopts John Curwen's thick, thin and small labelling (Curwen 2010), and he describes three registers for the male voice (*lower thick, upper thick* and *upper thin*) and five registers for the female voice (*lower thick, upper thick, lower thin, upper thin* and *small*).

Yet the terms used to label the vocal registers until today are abundant and author dependent, and most of the time their usage is ambiguous (Henrich 2006, see Figure 2.1.6). In 1963, a literature survey summary concerning the pitch range and labelling of voice registers (Mörner, Fransson and Fant 1963) mentions that

> "the only secure common denominator for defining a register is by means of its range on the musical scale.' Reasonable agreement is found on 'the average pitch of the boundaries between registers, i.e. the breaks or voice transitions." (Mörner, Fransson and Fant 1963)

Register is understood by many teachers, but also researchers as a perceptual phenomenon, in accordance with the homogeneity of the timbre. Johan Sundberg, the author of the seminal volume "The Science of the Singing Voice" (Sundberg 1987), describes a register as

> "... a phonation frequency range in which all the tones are perceived as being produced in a similar way and possess a similar voice timbre." (Sundberg 1987, p. 49)

He then goes on to study this perceptual phenomenon with all the observation and measurement methods available for vocal apparatus: spectral analysis for acoustic evaluations (including inverse filtering for vocal tract and vocal source decomposition), aerodynamic (oral and subglottal pressure, respiratory volume), electroglottography for vocal folds contact, video-fiberoptic observation for the supra-laryngeal pharynx configuration changes (strobolaryngoscopic and videokymographic) as well as dynamic real-time MRI imaging for the overall vocal apparatus modifications (Sundberg, Gramming and LoVetri 1991, Sundberg 1987, Echternach et al. 2008, Granqvist et al. 2003, Svec, Sundberg and Hertegård 2008, Cleveland, Sundberg and Prokop 2003).

Ingo Titze, another prominent singing voice researcher, defines registers as follows:

> "... the term register has been used to describe perceptually distinct regions of vocal quality that can be maintained over some ranges of pitch and loudness." (Titze 1994, p. 282)

Other professionals prefer a mixture of physiological (laryngeal) and perceptual (resonance) factors to delimit registers (Large 1972). Harry Hollien notes that a register must be operationally defined "1) perceptually, 2) acoustically, 3) physiologically and 4) aerodynamically" (Hollien 1974).

In the late 1970s, an international organisation composed of physicians, voice scientists, voice coaches and voice pathologists, a committee on vocal registers was formed in an attempt to clarify the notion of vocal registers and to find a consensual position among the international voice community (Henrich 2006). The committee has "accepted the notion that there probably are two sources for registers – the

larynx and the vocal tract" (Hollien 1983). However, this point seems to have raised a great debate among the committee members. "A substantial minority of the committee argued in favour of the source (of a voice register) being only laryngeal and that the other so-called register-like phenomena actually are some sort of quality/timbre events." (Hollien 1983)

Kob et al. 2011 summarises more recent studies which explore articulatory behaviours in the main singing-voice registers for both male and female operatic singers with MRI. The transition from modal to falsetto registers resulted in only minor modifications of vocal-tract shape, such as an elevation and tilting of the larynx and a lifted tongue dorsum. In comparison, the transition from chest to head register resulted in major modifications, such as a pharynx widening, lip and jaw openings, and increased jaw protrusion. Their results seem to comply with the notion of the middle passagio being a result of timbral (vocal tract) adjustments.

At the turn of the century the laryngeal mechanism of register transitions was studied by means of electroglottographic measurements (Henrich 2006, Roubeau, Henrich and Castellengo 2009b). Transitions from one mechanism to the other displayed EGG amplitude change, even if no pitch jumps were present and timbral differences were avoided by skilful singers. The authors detected three transitions and therefore four regions where a single laryngeal mechanism is evoked for vocal production. They numbered the regions – introducing the most neutral labelling, in accordance with the committee recommendations (Figure 2.1.7) - *M0* to *M3*, *M* standing for "mechanism".

M0, also called *vocal fry*, is characterised by short, thick, slack vocal folds, low activity of both thyroarytenoid (vocalis) and cricothyroid muscles (Hollien 1974, Hirano 1988). The closed phase of the vocal folds vibration cycle is longer than the open phase (Henrich 2006). Employing *M0* can help extend the vocal range of a low voice to even lower frequencies, like Russian basses. It can also occur when vocal folds are relaxed e.g. at the end of musical phrases, as can sometimes be heard in blues, rock or pop music.

M1 is the primary mechanism in the lower to mid range for male and female speakers and singers. In *M1* vocal folds are thick and vibrate over their whole length with a vertical phase difference (Vennard 1967). Thyroarythenoid (vocalis muscle) activity dominates (Hirano 1988). Closed phase is usually longer than open phase (Henrich 2006).

Compared to *M1*, in the laryngeal mechanism *M2* the vocal folds mass is reduced and there is no vertical phase difference (Vennard 1967). Cricothyroid muscle activ-

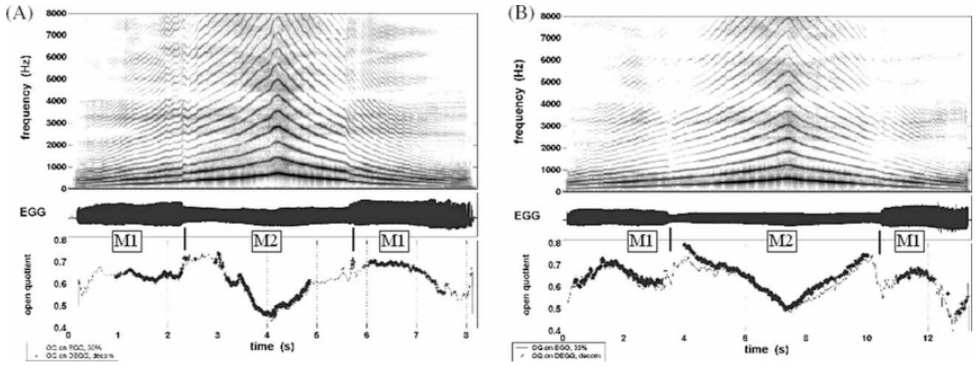


Figure 2.1.6: Vocal register transitions on a glissando sung by a counter tenor (A) with a pitch jump and a noticeable timbre change, (B) without any noticeable break. The top panel shows the time-frequency analysis, the middle panel the EGG signal, and the bottom panel the EGG measured open quotient. From Henrich 2006 with permission from Taylor & Francis.

ity dominates over the vocalis. Vocal folds are stretched thin and the edges (the ligament) are stiffer than the body (the vocalis muscle, Hirano 1988). The closed phase is shorter than the open phase of the vibration cycle (Henrich 2006). Also, the fundamental frequency dominates the spectrum (Sundberg 1987). *M2* is employed by female singers extensively in mid to high range. Male singers can produce a female-like, flute-like sound in their upper range using *M2*. For untrained singers it is often difficult to get vocal folds contact in *M2*, resulting in a hooty, breathy sound.

M3 enables singers to produce the highest pitches, sometimes called *whistle*, *flute* or *flageolet*. It has not been thoroughly studied. The vibration amplitude is reduced compared to *M2* and the vocal folds are thin and tense (Henrich 2006).

The range between *M1* and *M2* has been the most debated and the search for *"M1.5"* has been going for decades. Pushing the limits of register ranges and managing the transitions between registers are some of the main challenges in vocal technique and therefore a crucial part of a singer's education. Many singing teachers and researchers denote this range as a separate register with names like *mixed voice*, *middle register*, *voix mixte*, etc. Researchers questioned whether laryngeal mechanisms can be mixed to produce these perceptually distinguishable registers. Yet an additional laryngeal mechanism has not been found (Henrich 2006). It has been shown for Western lyrical singing that *voix mixte* is produced by male singers

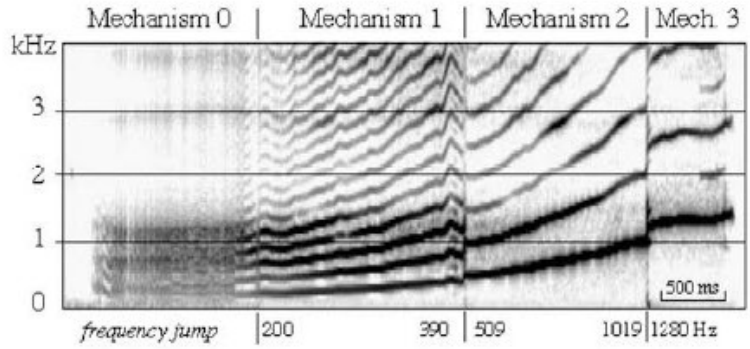


Figure 2.1.7: Sonogram of ascending vocal glissando performed by female subject showing successive use of laryngeal mechanisms M0-M4.(Roubeau, Henrich and Castellengo 2009b, p. 246). With Permission from Elsevier.

primarily in *M1*, while female singers usually employ *M2* (Castellengo, Chuberre and Henrich 2004, Lamesch et al. 2007). Studies on *belting* point to *M1* being used (Schutte and Miller 1993, Estill 1988, Henrich 2006). Yet an extensive use of twang also helps to mimic the *M1* sound and feel while in *M2* due to vocal tract shape influence on voice source (Sundberg and Thalén 2010).

2.1.3 Physiological building blocks – Jo Estill

Jo Estill (1921-2010), an American singer, singing teacher and voice researcher, is known for her revolutionary approach to singing education which, since 1970s, has changed the general discourse in this area. Based on her research and her understanding of vocal style and vocal physiology she introduced physiological "building blocks" as an instrument of singing voice manipulation as well as of vocal pedagogy. She aimed at isolated command of these building blocks, developed exercises to increase perception and control as well as to isolate activity of the building blocks. She then proceeded to using these isolated functions of the building blocks to combine them in various ways to create different vocal qualities (twang, sob, speech) and styles (opera, belt). Estill has had dedicated followers in most English-speaking countries and her influence on the singing teaching profession has been extensive.

2.1.3.1 Research, commerce, impact

After a carrier as a singer in the music industry Jo Estill received an MA in Music Education and was appointed Instructor in Voice in the Department of Otolaryngo-

Table 2.1: Pedagogues most regularly cited as influential in the development of teaching techniques for singing. From (Mitchell et al. 2003, p. 170). With permission from Taylor&Francis

Influence	Chapman	Estill	Miller
Pedagogues	1, 4, 7, 9, 15	1, 2, 3, 4, 7, 9, 10, 11, 13, 15	2, 4, 8, 9, 10, 13, 15
%	33	67	47

logy, at the Upstate Medical Center, Syracuse, NY, in 1972–1979. There she worked under Dr. Ray Colton, and Dr. David Brewer, two of the top voice researchers in the United States and began her pioneering research on her six voice qualities (Chapman 2011). Between 1980 and 1984, Jo enrolled in the City University of New York PhD graduate programme in Speech and Hearing. She completed all of her PhD course work and withdrew without submitting her dissertation. In 2004 Estill was awarded an Honorary Doctorate, Doctor of Letters (LittD), from the University of East Anglia[5].

She conducted research on vocal production physiology focusing on singing, employing measurement instruments such as EMG, electroglottography, voice signal analysis, X-rays of the phonating larynx, laryngeal fibre endoscopy, videostroboscopy and more (Estill 1988, Yanagisawa et al. 1989, Yanagisawa et al. 1991, Harris, Harris and Rubin 1998, Chapman 2011). She underwent very invasive measurement procedures as a subject to collect evidence for her ideas. Based on her research and experience she developed the Estill Voice Training system that was presented around the world and became influential in English-speaking countries. Many authors, singing teachers and performers have acknowledged Estill's influence in the development of their own work, including Gillyanne Kayes, author of "Singing and the Actor" (Kayes 2004), Deirdre Trundle (Trundle 2005), Donna Soto-Morettini (Soto-Morettini 2006), Lise Olson (Olson 2001) and many more. Mitchell et al. in their paper (2003) where they perform a content analysis of experts' pedagogical practices collected information on pedagogical influences of singing teachers; of their 15 participants 67% cited Estill as influential in the development of teaching techniques for singing (Table 2.1).

While Estill's research was published in scientific journals or reports, she approached her singing education innovation commercially and did not share any comprehensive summary of her system publicly. Her followers in singing educa-

[5]University of East Anglia Honorary Graduates: `https://portal.uea.ac.uk/graduationoffice/honorary-graduates`, last accessed 27/09/2017

tion have summarised and further developed her work and use their publications in commercial singing courses. Though some of them are also involved in scientific research, these commercial publications have not been subject to any external reviews and are not shared publicly[6].

The lack of public access and discussion of the Estill system has led to relatively little follow up research: Estill's articles are often hard to find, her booklets are not peer-reviewed. Her h-index of 13[7] shows that her work has not been cited a lot in spite of her huge influence. There seems to be little critiques of the Estill system either. Janice Chapman, an influential classical singing teacher, criticised the Estill Voice Training for not including breathing and abdominal support (Chapman 2011, p. 257). Jeannie LoVetri, a prominent US singing teacher who coined the term Contemporary Commercial Music, criticised the idea of deliberate manipulation of laryngeal structures in her popular blog[8], though she never directly addressed the Estill system in her critique. In July 2017 Gillyanne Kayes, the author of "Singing and the Actor" (2004) and one of the first singing teachers certified in the Estill Voice Training, gave a talk at the Association of Teachers of Singing conference entitled "After Estill"[9], in which she discussed new developments and changed practices in her own work. The video of the talk was watched thousands of times within a day (Gillyanne Kayes, personal communication) and generated a heated, sometimes offensive, Facebook discussion.

A transparent, public discussion of Estill's heritage is therefore necessary and timely. We welcome the new and long awaited publication of a volume on the Estill system based on her unpublished manuscripts (Steinhauer and McDonald Klimek 2017). Unfortunately, we have not been able to get hold of the book yet.

2.1.3.2 The Estill model

The Estill Voice Training™ teaches isolated control of individual anatomical structures within the voice production system (Estill et al. 2005a, p. 5).

The Estill Voice Model™ includes 13 physiological structures of the larynx and the vocal tract presented in Figures 2.1.8 and 2.1.9 (Figure 2.1.10 shows laryngeal

[6] A new book has been published recently (Benson 2017) promising public access to detailed discussions of the Estill system concepts for the first time. Unfortunately we have not been able to get hold of it.

[7] Calculated from Google Scholar citation counts on 27/09/2017

[8] e.g. `http://somaticvoicework.com/against-manipulation-in-the-throat/`, last accessed 27/09/2017

[9] `https://s3-eu-west-1.amazonaws.com/afterestilldrgillyannekayes/AFTERESTILLKeynoteAOTOSDrGillyanneKayes_player.html`, last accessed 27/09/2017

Figure 2.1.8: Physiological structures controlled in Estill Voice Training (Estill et al. 2005a, p. 5).

Figure 2.1.9: Overview of voice production structures and options in Estill Voice Training Level One (Estill et al. 2005a, p. 6).

structures in more detail). The assumption is that isolated control and manipulation can be taught for these structures and they can then be used as building blocks for a conscious construction of musical styles, introducing new sounds or increasing the variety of vocalising.

Vocal folds vibration mode Jo Estill in her approach refused to use the registration terminology. Since she based her method on physiology, she preferred to talk about the mechanics of vocal quality. She differentiates between modal and falsetto, in analogy to M1 and M2 mechanisms (see Section 2.1.2). Within the modal mechanism she talks about thin or thick vocal folds – a dichotomy defining a continuum of laryngeal settings in between. In her view trained singers can employ thick or thin vocal folds at any pitch (with a varying grad of difficulty). Moreover, vocalists can change gradually from one to the other (Estill et al. 2005a p. 43). Her approach to the vibratory patterns of vocal folds and the interaction between their layers with each other is based on M. Hirano's "body-cover model of fundamental frequency control" (Hirano 1974, Hirano and Kakita 1985, Titze 1994), where the deeper layers (the vocalis muscle and the ligament) form the body and the exterior layers (the superficial lamina propria and the epithelium) comprise the cover (Figure 2.1.11). She employs the Dynamical Systems Theory (Kelso 1997) to point to the reason for registration in order to dismiss the attractor states as being a given:

> "The complex mechanical interplay of body and cover as the length of the true vocal folds changes (via contraction of the thyroarytenoid and/or cricothyroid), together with the aerodynamic influence of the breath, results in different vibratory modes, gears or registers. Within this "dynamic system" are those attractor state vibratory modes that are most recognisable – for example, Modal/Speech and Loft/Falsetto.

Figure 2.1.10: Larynx rear and lateral view, exploded, after Sundberg 1987, p. 8.

Figure 2.1.11: Layered structure of vocal folds (Estill et al. 2005a, p.42).

> In Estill voice training, the challenge is to learn how to maintain the conditions that produce a given vibratory mode beyond the frequency boundary of its attractor state." (Estill et al. 2005a)

She introduces four vibratory modes: *Slack*, *Thick*, *Thin* and *Stiff* (Estill et al. 2005a, p. 44). *Slack* refers to the MO mechanism (Section 2.1.2). *Stiff* mode is characterised by a change of the vocal folds plane and little or no contact during the vibratory cycle:

> "May occur naturally at higher pitches where the vocal folds are elongated, pulled taut, and positioned slightly away from the midline. In some individuals this may be accomplished by cricothyroid activity; in others, the arytenoids may rock back, raising the back end of musculomembranous portion of the true vocal folds. Known as falsetto in Estill voice training, there is little or no contact during vibration, although the tone may or may not be breathy." (Estill et al. 2005a, p. 44)

The term *stiff* seemed misleading in this context since in physiology stiffing a muscle means activating it. She seems to believe that the vocalis muscle is active ("pulled taut") in this vibration mode. Other researchers disagree (e,g, Jillyanne Kayes, personal communication). We decided to use the more common and more neutral *falsetto* description.

Defining the *thick* vibratory mode Estill again refers to the register terminology, then reformulating the ambiguous language in physiological terms:

> "Known as the modal or speech register. The true vocal folds are relatively short, with some contraction/muscle tone in the vocalis muscle within the body. The cover is pliant, and the folds ripple (mucosal wave) from the lower border to the upper border of the fold edge, with a thick depth of contact through the cycle." (Estill et al. 2005a, p. 44)

Thick folds mucosal wave is illustrated in Figure 2.1.12.

Thin vocal folds are just that – due to elongation of the folds the edges become thinner leading to a shallower contact:

> "May occur naturally at higher pitches where the vocal folds are elongated, and/or during soft voicing. Less cover flexibility and/or less breath

Figure 2.1.12: Modal register/thick folds vibration mode: schematic of one cycle of the vocal folds as seen from frontal and coronal views, illustrating mucosal wave (Sundberg 1987, p.64).

Figure 2.1.13: Thyroid cartilage tilt. The cricothyroid muscles enable the movement around the cricothyroid joint, changing the position of the thyroid and cricoid cartilages relative to each other (Baken 2006, p. 71).

result in vibration without the lower-to-upper ripple, a thin or shallow depth of contact." (Estill et al. 2005a, p. 44).

Onset coordination The Estill model describes three ways to coordinate the beginning of exhalation with the movement of the vocal folds into position for vibration: in a *glottal* onset the vocal folds close before the exhalation begins; in the *aspirate* onset the exhalation begins before the closure of the vocal folds; the *smooth* onset is when the exhalation and the vocal folds movement begin simultaneously (Estill et al. 2005a, p. 26).

Thyroid and cricoid cartilage tilt The cricoid cartilage, shaped like a signet ring, forms the base of the larynx and sits on top of the trachea, to which it is connected via mucous membrane and connective tissue. The thyroid cartilage, having a shape of a shield, consists of two plates of cartilage fused at the front and wide apart at the back. The line where the plates are connected forms what is known as "the Adam's apple". Vocal folds originate from the vocal processes of arytenoid cartilages – small, pyramid shaped paired cartilages connected to the upper border of the cricoid cartilage lamina – and insert into thyroid cartilage at the line where the plates are fused (Figure 2.1.10).

Thyroid and cricoid cartilages are connected at the cricothyroid joint. They can rotate relative to each other along the axis of the cricothyroid joint. The space between the lower edge of the thyroid cartilage and the arch of the cricoid cartilage is called the cricothyroid space (CT space, Estill et al. 2005a, p. 51) or the cricothyroid visor. Opening or closing cricothyroid visor affects the length, thickness and therefore the vibration mode of the vocal folds as well as the fundamental frequency (pitch) of the sound. (Figure 2.1.13).

The Estill model separates the movement of the two cartilages and teaches to control them independently, also separate from the vocal folds body-cover state(Estill

(b)

Figure 2.1.14: Opening the cricothyroid space: a) cricoid tilting forwards, b) another strategy: thyroid tilting backwards (Estill et al. 2005a, p. 59 and 60).

et al. 2005a p. 55). For thyroid cartilage it introduces two states: a vertical thyroid and a tilted thyroid (Figure 2.1.13). Thyroid is vertical in quiet breathing. It tilts forward when the cricothyroid muscle is actively engaged, like in high meowing or whimpering. Thyroid cartilage tilt is said to reduce the interharmonic noise and be responsible to the perception of sweetness in the vocal tone (Estill et al. 2005a p. 56).

Titze indirectly supports this view. For him, tilting the thyroid is involved in register transition for Western classical singers:

> "Titze (1994) states that there are two theories about registration. The first involves the coordination between the cricothyroid muscles and the thyroarytenoid muscles. For example, singers gradually relax the thyroarytenoid (TA) muscles as the cricothyroid muscles (CT) gradually increase activation. The thyroid cartilage tilts forward, stretching the vocal folds, thereby increasing tension, for a smooth ascent of pitch." (Bateman 2010)

Also MRI investigations of speech found a forward rotation of the thyroid on higher pitches (Takano, Honda and Kinoshita 2004).

Cricoid cartilage tilt is related to shouting in the Estill model (Estill et al. 2005a, p. 59). The cricothyroid joint that allowed for closing of the CT space in thyroid cartilage tilt can also be rotated for an opening stretch of the cricothyroid membrane (Figure 2.1.14). It results in shortening and thickening the vocal folds which contributes to the loudness and defines the vocal quality of the human activity such as shouting or belting. Estill describes the physiological mechanism of the cricoid cartilage tilt based on the pulling force of the cricopharyngeus division of the inferior pharyngeal constrictor (Figure 2.1.15). An alternative strategy to shorten and thicken the vocal folds is suggested, where the thyroid cartilage tilts backwards instead (Figure 2.1.14), pulled by the middle constrictor or other muscles, thus opening up the CT space (Estill et al. 2005a, p. 60).

Figure 2.1.15: Pharyngeal constrictors (Estill et al. 2005a, p.94).

False vocal folds False vocal folds or ventricular folds are located above the vocal folds and are one of the three constrictors of the larynx (true vocal folds and AES being the other two). They close when the body prepares for a strenuous activity such as lifting heavy objects, defecating or childbirth, or when you prepare to fight, flee or scream (Estill et al. 2005a, p. 33). The three conditions of the false vocal folds in the Estill model are:

mid: of comfortable speaking/singing,

constricted: moved inwards, and

retracted: moved outwards (Estill et al. 2005a, p. 35).

While the constriction of the false vocal folds is observed in straining, the opening of the false vocal folds is associated with laughter and crying. The intrinsic muscles of the larynx are thought to close both the true and the false vocal folds together. The muscles that allow the false vocal folds to be opened and closed independently of the true vocal folds have not been identified, but may include extrinsic muscles of the larynx (Estill et al. 2005a, p. 34, see Figure 2.1.18).

AES Numerous studies found that supraglottic compression contributes to volume, vocal brilliance and efficient voice production (Bartholomew 1934, Rothenberg 1981, Sundberg 1987, Hirano 1988 Yanagisawa et al. 1989, Titze 2008, Guzman et al. 2015). Manuel Garcia in his report to the Royal Society of London in 1855 noted that epilarynx greatly affects the quality of the voice:

> "... by its contraction it gives brilliancy to it and by its widening volume."
> (García 1855)

Jo Estill introduced the narrowing or widening of the aryepiglottic sphincter as the physiological building block responsible for the epilaryngeal compression (Figure 2.1.16a). *Aryepiglottic sphincter* (AES) is the upper of the three laryngeal sphincters alongside the vocal folds and the ventricular folds (false vocal folds, Pressman 1954). The constriction of AES is caused by the approximation of the tubercle of the epiglottis (anterior), aryepiglottic folds (lateral), and arytenoids (posterior, Figures 2.1.16b and 2.1.17).

(b) narrowing muscles

Figure 2.1.16: AES (Estill et al. 2005a, p. 87).

Figure 2.1.17: Posterior view of larynx showing aryepiglottic and oblique arytenoid muscles (Kayes 2004, p. 111).

In the Estill model AES constriction is the main building block of the vocal quality of *twang* (Colton and Estill 1981). *Twang* is a piercing, often nasalised sound we are used to in Country music, in many Northern American accents, in children's voices ringing around the schoolyard, in belting sounds of musical theatre. To experience this sound quality, one is usually asked to utter a loud "Miaow!" or to sing the playground taunt: "Nyae – nyae nyae nyae – nyae!" (Estill et al. 2005a, p. 87). In the Estill model AES narrowing (Figure 2.1.16) often seems to be used synonymously with the term *twang* because it is the primary physiological feature of this voice quality (which also displays a high larynx, a high tongue, thin vocal folds and a tilted thyroid, Estill et al. 2005b, p. 43). Estill distinguishes between the nasal and oral twang: AES constriction can be produced with the nasal channel open or closed (Estill et al. 2005b, pp. 41ff).

In their 1989 paper Yanagisawa, Estill et al. observed AES narrowing by means of fiberoptic videolaryngoscopy in their 5 subjects in three voice qualities: *twang*, *belting* and *opera*. They likened the vocal gesture associated with epilaryngeal constriction to arrested swallowing (Yanagisawa et al. 1989).

A number of voice qualities have been shown to rely on activity in the epilaryngeal area in speech and in voice pathology. While breathy voice implies an open sphincter, 'creak' (or vocal fry) results from a constricted sphincter (Esling, Harris and Romero 2003). Other qualities such as whisper and harsh voice are also a function of laryngeal sphinctering (Moisik and Esling 2011).

Voice qualities are important not only in phonetics but also in vocal health and vocal education. Catherine Sadolin's Complete Vocal Technique (Sadolin 2000) is famous for approaching vocal effects such as growl as something that can be taught and used safely as opposed to voice pathology or inefficiency needing correction. Growl is produced through co-vibration of vocal folds and aryepiglottic folds which results in subharmonic oscillation. (Moisik, Esling and Crevier-Buchman

2010, Sakakibara et al. 2004a).

Linguists have been interested in the constriction of the epilarynx in relation to consonants produced in this part of the vocal tract, for example in Arabic (Zeroual, Esling and Crevier-Buchman 2008) and in other languages. In particular, they studied the role of the aryepiglottic sphincter in the glottal and the epiglottal stop (Esling, Harris and Romero 2003). Both gestures involve partial ventricular fold adduction alongside AES constriction.

Estill mentions that AES narrowing can lead to false vocal folds (ventricular folds) constriction (Estill et al. 2005a, p. 88). In the spirit of her system, Estill teaches to separate the two constrictions in order to control them independently. More recent studies confirmed that experienced singers in power-intensive styles did separate the two constrictive behaviours: Guzman et al. showed that anterior-posterior laryngeal compression (approximation of the tubercle of the epiglottis and arytenoids) was about 10 times higher than the medial compression (ventricular folds approximation) in Rock, Pop and Jazz singers (Guzman et al. 2015). Similar results for Western classical singers were given in Mayerhoff et al. 2014.

Twang is characterised by the bright, "ringy" sound quality which is believed to result from a positive slope of the spectrum (Lichte 1941). Helmholtz (1877) described a "ringing, clear tinkling as of little bells," in the presence of the higher partials from 2,640-3,168 Hz. Yanagisawa et al. 1989 found that narrowing of the AES produces a peak in the spectrum in the region between 2000 and 4000 Hz. This bandwidth corresponds to the resonant frequency of the outer ear canal, making the sound perceptually louder and more "ringy" (Bartholomew 1934, Yost 1994). In particular, for fundamental frequencies below 1000 Hz, there can be a 15-30 dB advantage in sound transmitted to the middle ear without an increase in vocal effort (Yanagisawa et al. 1989).

Sundberg found that when the cross-sectional area in the pharynx is at least six times wider than that of the laryngeal tube opening, the epilaryngeal tube is acoustically unlinked from the rest of the vocal tract acting as a separate resonator. Therefore, an extra formant is added to the vocal tract transfer function contributing to the generation of the "singer's formant" – a concentration of acoustic energy around 3000 Hz (Sundberg 1987, chapter 5). Rothenberg hypothesised that in this case the laryngeal tube is no longer coupled with the supraglottic vocal tract, but interacts more strongly with the vocal folds (Rothenberg 1981). Titze agrees that the narrowing of the epilarynx produces the non-linear acoustic coupling between the glottis and the laryngeal tube. In this coupling the glottal impedance is adjus-

Figure 2.1.18: Extrinsic muscles of the larynx responsible for its movement (Estill et al. 2005a, p. 65).

ted to be comparable to the vocal tract input impedance making the glottal flow highly dependent on the acoustic pressure in the supraglottal vocal tract (Titze 2008). The inertive vocal tract enhances vocal fold vibration because the supraglottal pressure driving the airflow is synchronous with the velocity of the vocal folds. By means of AES narrowing an inertive vocal tract is formed which facilitates vocal folds oscillation and increases the amplitude of the singer's/speaker's formant (Lombard and Steinhauer 2007). This theory has been supported by empirical evidence from several studies where increase in perceived loudness and brightness was found (Yanagisawa et al. 1989, Sundberg and Thalén 2010, Mayerhoff et al. 2014, Guzman et al. 2015).

The physiological mechanism of aryepiglottic sphinctering is controversial (Moisik 2008). The supraglottal constriction is considered to be caused by an activity of the aryepiglottic muscle, however, several researchers do not support this view and empirical data is not conclusive (Sakakibara et al. 2004b). It is likely that the burden of laryngeal sphincter during swallowing and constricted articulations is a result of synergistic activity of the muscles (Moisik 2008).

Larynx height We can change the position of our larynx and through that the length of our vocal tract as well as its resonance frequencies (the formants, Sundberg 1987). Some flexibility in laryngeal movement is imperative for healthy singing. Singing styles and traditions differ in the amount of this flexibility they require. While Western classical singing encourages a stable and for a large part of the range a lowered larynx, Ottoman classical tradition as well as some Western pop music repertoires require a large palette of laryngeal positions.

The muscles responsible for the larynx movement are those attached to the hyoid bone: infrahyoid muscles lower the larynx (with the exception of thyrohyoid which can act as an elevator) and suprahyoid muscles raise it (see Estill et al. 2005a, p.65). Pharyngeal constrictors also play a role in raising the larynx. The muscles of the tongue and velum provide support (Figure 2.1.18).

The neutral position of the larynx is the one at rest, during quiet breathing (no phonation). Low larynx occurs on low pitches or in preparation to sob. High larynx is related to high pitch or to a preparation to scream or squeal (Estill et al. 2005a,

p.66).

Changes in larynx position can be measured in a singer in a non-invasive manner with a good precision by means of a laryngoaltimeter (Pehlivan and Denizoglu 2009), which compares the signals obtained from two matched microphones placed on the suprasternal notch and supraglottic region.

The problem with determining the larynx height is not so much the objective measurement as the fact that larynx position changes with pitch. Therefore, in singing, each pitch will have its own neutral position for a given singer, the one that the singer perceives as the most comfortable or a listening expert would rate as the most appropriate. Both judgements are intrinsically subjective; the former is defined by the singer's habits; the latter by the expert's background and preference. Yet singing teachers pass the judgements about larynx position routinely in relation to pitch – it is the one measure which matters to singing as opposed to the overall larynx height which reflects the pitch in the first place. This contradiction makes the definition of larynx height ambiguous.

It is reported that intensity level affects larynx height being lower in loud phonation and higher during soft productions. Guzman (2015) hypothesises that "laryngeal lowering during loud phonation acts as a protecting factor". Sundberg claimed larynx to be higher in belting. Shortening of the vocal tract in this way enables the singers to track the first harmonic by the second formant (Schutte and Miller 1993).

Velum Velum or soft palate is a posterior continuation of the hard palate of the mouth, together they build the superior wall of the oropharynx (Figure 2.1.2). Velum is movable and consists of muscle fibres and mucous membrane. It functions like a door opening and closing the velopharyngeal port which connects the oropharynx (mouth cavity) with the nasopharynx (nasal passages). It closes off the nasal airway during swallowing, while in sneezing it directs a portion of the excreted substance to the mouth to protect the nasal passages. During speech and singing lowered velum results in nasal sounds while raised velum produces orally released vowels and consonants. Closure of the velopharyngeal opening is also required to create the intra-oral pressure needed to form the plosive and fricative consonants.

The Estill model operates with three velum positions: low – nasal sounds, mid – nasalised and high – oral vocalisation (Estill et al. 2005a, p. 72). The textbook speaks about oral, nasal and mixed resonance for high, low and mid velum, e.g. "the nasal consonants are resonated in the nasal passages only" (Estill et al. 2005a, p. 72). This is not quite accurate. For nasal consonants (nasal occlusives) [m], [n] and

[ŋ], which are among the most common sounds cross-linguistically, the air escapes only through the nose, but the oral cavity still acts like a resonance chamber. When the velum is high and the nasal passages are separated, the mouth cavity is in fact the main resonator, as is the case with most English vowels. With mid velum, the air escapes though both mouth and nose, producing nasalised sounds, typical in some languages such as French.

Whether a specific nasal resonance plays an important role in vocal sound is questionable. Sundberg et al (2007) found that an idealised model of nasal passages of a baritone singer, constructed of iron tubes, produced a resonance. But when the model was made of epoxy, the resonance was heavily damped, particularly when the maxillary sinuses were included in the nasal system (Sundberg et al. 2007). Yet opening the velopharyngeal port affects higher harmonics produced in the mouth (Yanagisawa, Kmucha and Estill 1990). The studies are contradictory about the exact effect of lowering the velum on the spectrum. Estill textbook points out a general dampening of the higher harmonics in an isolated exercise (Estill et al. 2005a, p. 73). Sundberg concludes that singers seem to be able to enhance higher spectrum partials by a careful tuning of a velopharyngeal opening (Sundberg et al. 2007). The coupling between the oropharynx and the nasopharynx may indeed be used a lot by tenors or female belters to achieve high pitches (Cross 2007, cited in Garnier et al. 2007a), while Western classical singers often nasalise the vowel [a] (Birch et al. 2002, Sundberg et al. 2007).

Tongue The tongue is a large muscular organ in the oral cavity that plays a major role in articulation and in shaping resonance in speech and singing (Figure 2.1.19).

Vowels differ from each other mainly in the position of the tongue. In phonetics vowels are commonly plotted as a two-dimensional chart with dimensions representing vowel closeness and vowels backness or the first and the second formant (IPA chart). The first formant is associated with the vertical position of the tongue, while the second formant is particularly sensitive to the tongue shape, the frontness/backness of the tongue body (Sundberg 1987, chapter 5). The tip of the tongue when advanced or retracted will raise or lower the third formant. When the tongue body is pulled back it constricts the oral-pharyngeal resonating space and will raise the first formant (Sundberg 1987). In pedagogical texts this action may be referred to as 'tongue root tension' (Chapman 2011). The epilaryngeal tube physically links the larynx with the tongue, therefore an excessive tension of the tongue directly affects the freedom of movement of the larynx. The singers have to compromise

Figure 2.1.19: Parts of the tongue (Estill et al. 2005a, p. 79).

between the clarity of phonetic articulation on the one hand and vocal resonance such as clustering formants or adjusting formants to harmonics on the other hand, and tongue shape and position is crucial in this respect. For example, Western classical singers tend to curl the tip of the tongue to raise the third formant: the fifth formant is lowered so that the third, forth and fifth formants build a cluster generating an energy peak around 3000 Hz, known as *the singer's formant*, which helps to carry the sound of the singer's voice above the orchestra without external amplification (Sundberg 1987, chapter 5).

Estill emphasises that activity in the tongue tip and blade can be isolated from activity in the dorsum and root (Estill et al. 2005a, p. 79). There are four states of the tongue that are given in the Estill model:

low: a low flat tongue producing a dark tone

mid: the tongue dorsum is in speech position for general English language

high: the dorsum and root of the tongue are lifted

compressed: the tongue tip is curled up, the back of the tongue is pushed forward.

Articulation: jaw and lips In the Estill model the jaw position is described in four states: forward (lower teeth ahead of the upper teeth), mid, back (over-bite) and drop (the extreme opening, Estill et al. 2005a, p. 95). Noticeably, apart from the extreme drop, the width of jaw opening is not directly rated in the system. This is presumably to separate phonetic articulation from vocal adjustments. Like the tongue (see Section 2.1.3.2) the jaw opening can change with every sound we pronounce. At the same time the jaw position can be adjusted to influence vocal resonance, e.g. protruded jaw and lips would lengthen the vocal tract. For example in the villages of Northern Russia traditional singers sing with their mouths barely open, while in western regions of Russia singers often use protruded, wide open jaw, though the exact opening would depend on the vowel. It has to be noted though that there are still both dimensions in the Estill's description of the jaw position: forward/mid/back and neutral height vs drop. The textbook explicitly mentions that the jaw can be dropped from the forward position as well as from the mid position (Estill et al. 2005a, p. 97).

Lips position is described in three states: protrude, mid, spread (Estill et al. 2005a, p. 100). With lip protrusion, the overall length of the vocal tract is increased, supporting resonance of lower frequencies/harmonics; with lip spreading it is shortened, supporting resonance of higher frequencies/harmonics (Estill et al. 2005a, p. 99). Again, ranking the actual lip opening is avoided since it is determined by phonetics in the first place. Protrusion and spread are also highly dependent on the vowel and can follow each other on a millisecond scale. Yet if the lip form for each vowel is consistent and vowels constitute the main part of the duration of a vocal expression, the lip form would make an important contribution to the vocal resonance and timbre.

Posture and support Head and neck anchoring is introduced in the Estill model as "bracing" of skeletal structures; the large muscles (sternocleidomastoids at the sides of the neck, soft palate muscles raising the sail and the occipital region) are engaged. The function of the anchoring is to allow small muscles that control the vocal folds to work less hard and to be able to fine-tune their adjustments within a stable framework (Estill et al. 2005a, pp. 105-106).

Torso anchoring also has a support function stabilising the spine and rib cage. Pectoralis major of the rib cage, latissimus dorsi at the back and quadratus lumborum at the lower back are engaged (Estill et al. 2005a, pp. 111-112).

2.2 Ontology of vocal production

We combined in our ontology the voice source descriptors studied extensively by Johan Sundberg (Section 2.1.1), the classical registration terms (Section 2.1.2) and physiological building blocks introduced by Jo Estill (Section 2.1.3.2). Our descriptors therefore cover the voice source aerodynamics, laryngeal and vocal tract physiology. In this section we present our considerations about each of the descriptors, their measurement scales and the corresponding metrics.

Table 2.2: Ontology of vocal production

descriptors	**physiological dimensions**	**range**	**scale**	**metrics**
subglottal pressure	subglottal pressure	low to high	5-point	interval

Table 2.2 Ontology of vocal production continued...

descriptors	physiological dimensions	range	scale	metric
transglottal airflow	transglottal airflow	low to high	5-point	interval
phonation	phonation breathy	present/absent	2-point	nominal
	phonation pressed	present/absent	2-point	nominal
	phonation neutral	present/absent	2-point	nominal
	phonation flow	present/absent	2-point	nominal
register	vocal fry	present/absent	2-point	nominal
	chest	present/absent	2-point	nominal
	head	present/absent	2-point	nominal
	falsetto	present/absent	2-point	nominal
	flute	present/absent	2-point	nominal
vocal folds vibration mode	vocal folds vibration mode thick to thin	thick/ mixed thicker/ mixed/ mixed thinner/ thin	9-point, NA	interval
onset	aspirate	absent/ occasional/ often	3-point	interval
	smooth	absent/ occasional/ often	3-point	interval
	glottal	absent/ occasional/ often	3-point	interval
thyroid cartilage tilt	thyroid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
cricoid cartilage tilt	cricoid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval

Table 2.2 Ontology of vocal production continued...

descriptors	physiological dimensions	range	scale	metric
false vocal folds	false vocal folds	retracted/ mid/ constricted	5-point	interval
aryepiglottic sphincter	aryepiglottic sphincter	wide to narrow	5-point	interval
larynx height	larynx height	very low to very high	9-point	interval
velum	velum	low to high	5-point	interval
t**ongue**	tongue height	low to high	5-point	interval
	tongue compression	present/absent	2-point	nominal
jaw	jaw position	back/mid/forward	5-point	interval
	opening	minimal/ mid/ drop	5-point	interval
lips	protrusion	no/ slight/ strong	5-point	interval
	spread	no/ slight/ strong	5-point	interval
head/neck	head/neck anchoring	relaxed/anchored	2-point	nominal
torso	torso anchoring	relaxed/anchored	2-point	nominal

Voice source We summarised Johan Sundberg's work on vocal source in Section 2.1.1. Because the relationship between phonation modes, subglottal pressure and transglottal airflow is not as straightforward as of direct proportionality (see Section 2.1.1.1), including all vocal source descriptors into the ontology is justified. In particular breathy and pressed phonation have a history in other research fields; participants with the vocal health background might interpret them according to their clinical practice, while subglottal pressure and transglottal airflow are unambiguous physical characteristics.

The scales for subglottal pressure and transglottal airflow were chosen so as to correspond to real-life differentiation for practical purposes analogous to Estill scales: choosing between low, mid or high with the option to choose two adjacent states.e.g. low to mid.

For phonation modes four nominal dimensions reflect presence or absence of each mode (including Sundberg's *flow* phonation, see Section 2.1.1.1). This is necessary because the four phonation modes are not ordered. Presenting them as four independent dimensions would allow to indicate the presence of any combination of them.

Register/part of register range The laryngeal mechanisms *M0-M4* (Section 2.1.2) have the advantage of objectivity and, to some extent, measurability – the characteristics we seek for our ontological descriptors. Yet in the singing voice community they are not common terms, the definition, numbers and labelling of registers are still a matter of debate, and they continue to vary among authors. Indeed, perceptual vocal registers have an acoustical and perceptual reality for singers, which cannot be ignored (Henrich 2006).

We have included registers into our ontology using the most common terminology of *vocal fry*, *chest*, *head*, *falsetto* and *flute*.

Vocal folds vibration mode Because the thickness of vocal folds can be changed gradually (see Section 2.1.3.2) we have introduced a 9-point interval descriptor to our ontology called *vocal folds vibration mode thick to thin* to allow for that gradual change. E.g. if stands for *thick folds*, then 2 would denote *thick to mixed thicker* and 3 *mixed thicker folds*.

This descriptor only applies to the *M1* (modal) laryngeal mechanism. Therefore an *NA* value is necessary to indicate absence of *M1*, that would not affect the linear scale.

Onset The Estill model introduces three kinds of onsets: aspirate, smooth and glottal (Section 2.1.3.2). In a fragment of singing all kinds of onsets can be present and the amount of these onsets is informative. We therefore suggest an interval scale for the presence of each of the onsets.

This descriptor differs from others in that onsets are rare events. In a fragment of singing there may be no onsets. While all other descriptors will have a value for each singing fragment, onset may have none. In this case, and only in this case all

three variables will have the numerical value corresponding to the absence of this kind of onsets. Therefore, no extra *NA* value is necessary.

Thyroid and cricoid cartilages, false vocal folds, AES We chose a 5-point scale for other laryngeal structures such as thyroid cartilage tilt, cricoid cartilage tilt (Section 2.1.3.2), false vocal folds (Section 2.1.3.2) and aryepiglottic sphincter (Section 2.1.3.2) to enable a nuanced rating, as we have witnessed it in singing teachers' practice.

Vocal tract structures: larynx height and velum Both velum (Section 2.1.3.2) and larynx height (Section 2.1.3.2) are represented by three states in the Estill model: low, mid, high. We introduced a 5-point scale for the velum to allow for the combined states like "mid to high", which are very common in singing teachers' practice.

For the larynx height we decided on an even more fine-grained scale knowing from experience that teachers can sometimes give very detailed judgements about the larynx movements. Thus if 1 stands for *very low* larynx, then 2 expresses *very low to low*, 3 means *low*, etc.

Articulation Articulation descriptors tongue (Section 2.1.3.2), jaw and lips (Section 2.1.3.2) will change with each phonetic sound.

We included the low/mid/high position of the tongue as an interval dimension and added a nominal dimension to record the presence or absence of tongue compression (curling up the tip).

The jaw descriptor has two independent dimensions, as it can be dropped from a forward position as well as from a mid position. We suggest a finer quantification than in the Estill model: a 5-point interval scale for both anterior/posterior position and opening.

Various lip configurations can follow each other in quick succession (unlike the jaw) in any singing fragment. Therefore we represent the lips form with two dimensions: protruded and spread, indicating their presence (absence of both stands for the mid form) and the amount of deviation from the mid form.

Posture and support Nominal scales for both head/neck and torso anchoring are justified since the muscles described in Section 2.1.3.2 are either engaged or not. These two descriptors are usually determined visually by singing teachers, though

head and neck anchoring sometimes does have clear acoustic outcomes. Mostly they would be very hard to judge from audio recordings without any visual information.

2.3 Deconstructing Cantometrics parameters

We have discussed in detail physiological, aerodynamic and acoustic descriptors of vocal production and have compiled an ontology (Table 2.2). Now we return to the Cantometrics parameters and investigate them with the help of our more formal vocabulary.

Cantometrics musical parameter system was developed by Alan Lomax and his collaborators in order to describe and compare singing performance practice across societies. The system comprises 36 parameters quantifying various aspects of performance practice, one of which is vocal production. While it seems plausible that parameters like *social organisation of the singing group* may reflect other social organisation patterns of the society, it is much less intuitive that the preferred vocal timbre would tell us something about how that society works. Yet Lomax found a link between all of his parameters and societal traits.

Cantometrics parameters describing vocal production include: vocal width, rasp, nasality, volume, accent, glottal shake, vocal pitch. There are other parameters related to vocal production such as phrase length, melodic range, interval size, embellishment, tempo, tremolo, enunciation. At the outset we were particularly interested in the descriptor Lomax called *vocal width* for two reasons: a) there were obvious ambiguities in its definition and b) its alleged correlation with subordination of women made its investigation relevant beyond MIR and Cantometrics.

Lomax chose his parameters with the view that they should be easily understood by (any) raters after a short training. In his textbook he often does not give any significant definitions expecting the names to be self-explaining (e.g. volume or tempo). Sometimes one or two synonyms of the parameter name appear, otherwise Lomax prefers to add in some of his findings on the relationships to other cultural traits. Here is an example of an explanation text on the parameter called *vocal pitch*:

> "Register or pitch level seems to be another function of energy. Frequency of high register is associated with complex, exploitative productive systems and especially with a high calorie food intake ... . Much low register is characteristic of non-intensive agriculturalists." (Lomax 1977, p. 118)

The scales on which parameters are rated can give more information, e.g. *accent* very forceful to very relaxed. Then raters are given musical examples for each point of the scale.

In contrast, *vocal width* does have some sort of verbal definition:

> "The measure concerns the contrast between the voices which sound mellow, relaxed and richly resonant (we call them wide) and the voices which sound tense, pinched, and restricted in resonance (which we call narrow). Many singing styles can be characterised as having one or the other; in some rare cases both may occur; and many ways of vocalising (like everyday American speech) are neutral in width – these we call mid, singers with a "speech" tone." (Lomax 1977, p. 125)

The oppositions he uses are narrow/wide, tense/relaxed, richly resonant/restricted in resonance. He often names open throat as a characteristic of wide singing, therefore we can conclude that his notion of width/narrowness is related to the physiological state of the vocal tract. Yet what exactly should be wide/narrow? Is it jaw opening, lips form, tongue position? Is it the epilaryngeal opening dimensions? Anterior/posterior (AES, middle constrictor) or lateral (ventricular folds)? Even in speech phoneticians name several levels of constriction of the vocal tract (Moisik and Esling 2011).

The second opposition – tense/relaxed – is not less ambiguous. As we have seen in the previous sections there are many different groups of muscles that can be activated in the vocal tract. In the field of vocal health strain and hyperfunction could be related to Lomax's tension. There is evidence that these descriptors are notorious for their association with poor listener reliability and agreement (Kreiman et al. 1993, Oates 2009). Restricted resonance is not self-explaining either: who is more resonant, an opera singer or a belter? A classically trained vocalist with a high fundamental, or a folk singer with a high first formant?

From the physiological/acoustic point of view each of the oppositions requires a more formal definition. There is another question to be raised here – why are they all piled into one parameter? Lomax considered these oppositions to be closely related, almost interchangeable. He occasionally called the *vocal width* parameter *vocal tension*, and Victor Grauer, ethnomusicologist and Lomax's main collaborator on the Cantometrics project, referred to it as *vocal tension* in his writings (Grauer 2006a). Unfortunately, Lomax did not give any evidence why this would be the case. He consulted a medical voice specialist who, it seems, was the source of this

claim.

The Estill model provides us with a number of good counterexamples where these three oppositions are independent of each other. *Falsetto* in Estill's terms, which is the flute-like, vibrato-free quality found in "pure" children's voices and in breathy untrained vocalisation in the upper range, does not involve AES narrowing. It is in fact a very relaxed sound, apart from the high airflow, which cannot be forced. Yet there is nothing "wide" about it: all the vocal tract structures are in a mid, relaxed state, including the ventricular folds, larynx, the soft palate sail (head and neck anchoring in Estill's vocabulary), tongue, jaw, lips (Estill et al. 2005b, p. 21ff).

Another vocal quality prominent in the Estill system is called *Sob*. It is a soft and dark, emotionally intense in its pure form, mourning of an adult. All vocal tract structures are maximally expanded; the silent, suppressed sobbing requires intense muscle effort. The larynx is extremely low; false vocal folds completely retracted; head and neck anchoring at its peak, raising and tensioning the soft palate sail, AES is wide. This is probably the widest possible configuration of the vocal tract. At the same time, it is hardly relaxed; in contrast, the effort values to expand the vocal tract are among the highest.

Lomax seems to associate classical Western operatic sound with the wide and relaxed part of the spectrum of his *vocal width* parameter, especially mentioning the resonate quality (Lomax 1977). Estill sees the *Opera* vocal quality as a balanced mixture of *Speech* (mid, relaxed), *Sob* (wide, effortful) and *Twang* (narrow). In particular, she teaches to create the *Sob* quality with the low larynx, retracted ventricular folds and active head and neck anchoring, but holding to the narrow AES like in *Twang* at the same time. Velum is high, the tongue is compressed. Therefore, from the physiological point of view, this configuration is widened in some dimensions and narrowed in others, it demonstrates high levels of effort. Contemporary literature confirms AES narrowing in various singing styles including Western opera. Epilaryngeal narrowing is associated with semi-occluded vocal production which has a number of benefits (Titze 2006). Of 15 singing teachers in Mitchell et al. 2003 experiment only one explicitly associated open throat with relaxation and only five mentioned resonance/formants.

We summarised these three examples in Table 2.3. If Lomax's assumption of the dependency between width and tension were true we would see this correlation in the last three columns: e.g. wide vocal tract would correspond to low effort levels and high resonance and narrow vocal tract to high effort levels and low resonance. For all three examples this is not the case. Instead, we see a rather fuzzy picture. It

Table 2.3: Vocal qualities of *Falsetto*, *Sob* and *Opera* from the Estill system and their relationship with Cantometrics *vocal width* oppositions narrow/wide, tense/relaxed and richly resonant/restricted in resonance.

Estill vocal qualities	AES	larynx	head and neck	velum	tongue	**vocal tract width**	**effort levels**	**Resonance**
Falsetto	wide	mid	relaxed	low	mid	**mid**	**low**	**low**
Sob	wide	low	anchored	high	low	**wide**	**high**	**low**
Opera	narrow	low	anchored	high	compressed	**partly narrow, partly wide**	**high**	**high**

is this fuzziness that strongly indicates that no simple construct or definition will do justice to the complex phenomenon (or, rather, phenomena) which Lomax intended to capture in his *vocal width* parameter.

It is important to remember that Lomax was not a singer or a voice specialist and relied on the judgement of his medical consultant, whose background presumably only included Western musical traditions. There had probably never been a collaboration between a voice clinician and an ethnomusicologist before – Lomax's pioneering work has to be given credit.

2.4 Vocal width/vocal tension from the viewpoint of our ontology

In Section 2.3 we identified three components (dimensions) of the Cantometrics parameter *vocal width*. We demonstrated that these components are not equivalent. Let us now explore which descriptors from our ontology can possibly contribute to the perception of these components (see Table 2.4).

On the voice source side, it could be argued that pressed phonation can be associated with increased tension in the voice. In vocal health it is synonym to hyperfunction or extensive tension (Section 2.2). A question mark needs to be added here, due to the fact that Cantometrics studied healthy singers. In contrast, flow phonation was introduced by Sundberg to capture the vocal production used by Western classical singers; Lomax associated their technique with open, relaxed,

resonant sound. Neutral phonation could be positioned in the middle according to this logic. Breathy phonation is, on the one hand, rather relaxed, but on the other hand, certainly restricted in resonance. It is another example of vocal production which does not fit the Cantometrics vocal width scale (compare Section 2.3).

Register or vocal folds vibration mode do not seem to correspond closely with any of the *vocal width* components. Glottal onsets and constricted false vocal folds are associated with strain, aspirate onsets and retracted false vocal folds – with relaxation.

AES narrowing is a mechanism of an hypopharyngeal constriction, affecting directly the width of the hypopharyngeal opening. Narrowing AES is claimed to add twang, or ring, or brightness to the voice (Section 2.1.3.2). This is a resonance phenomenon, yet there is no clear correspondence to Lomax's dichotomy of richly resonant/restricted in resonance.

The position of the larynx determines the length of the vocal tract filter: the lower the larynx, the longer the filter (Section 2.1.3.2). While it does not literally affect vocal tract width, it is safe to suggest that Lomax considered a low larynx to be a characteristic of a wide vocalisation: Western classical singing which he considered the widest and most relaxed employs a low larynx.

High larynx is often associated by health professionals and singing teachers with a pressed, strained or unhealthy production. It could be argued that raising the larynx in a controlled way to add tension or to create an impression of tension in singing. Whether this is always the case needs further investigation.

Lowered velum could be associated with a narrowed mouth cavity, raised velum with widening it. In terms of resonance, lowered velum dampens the higher partials (Section 2.1.3.2).

A low tongue is consistent with a wider mouth cavity and a high tongue narrows it. Compressing the tongue like in classical singing to raise the third formant could be seen as a contribution to spectral richness (Section 2.1.3.2).

The relationships listed here are hypothetical, based on the analysis of our descriptors presented in this chapter. A valid examination of these relationships will require a dataset of singing recordings with annotations on vocal width and on each of our descriptors. Then a statistical evaluation can be conducted – we describe its methodology in Section 4.7.

Table 2.4: Ontology dimensions vs *vocal width* components

ontology descriptors	**ontology dimensions**	**width**		**tension**		**resonance**	
		wide	narrow	relaxed	tense	resonant	restricted
phonation	phonation breathy			✓			✓
	phonation pressed				✓		
	phonation neutral						
	phonation flow	✓		✓		✓	
onset	aspirate			✓			
	glottal				✓		
false vocal folds	constricted				✓		
	retracted			✓			
AES	narrow		✓				
	wide	✓					
larynx height	low	✓					
	high				✓		
velum	low		✓				✓
	high	✓					
tongue	low	✓					
	high		✓				
	compressed					✓	

2.5 Summary

In this chapter we discussed physiological descriptors of vocal production from Johan Sundberg's work, from classical theory of vocal registration as well as from the Estill model. We described in detail the physiology and the vocal function of the components and showed how their states were represented in our ontology summarised in Table 2.2. We concluded with deconstructing Cantometrics parameters based on the terms and concepts presented in the previous sections and hypothesising about physiological characteristics which contribute to their perception.

In the following chapters two approaches to automatic classification of our ontological descriptors are considered. Chapter 3 is dedicated to an incremental approach: we present a proof-of-concept experiment on automatic detection of phonation modes for sustained sung vowels. Generalisation to more varied datasets is discussed in Section 3.4. Chapter 4 lays out a methodology for an integrated approach, when reliable annotations are generated for the original dataset with all its variability.

3 Automatic detection of phonation modes

In the previous chapter we compiled an ontology of vocal production (Table 2.2) which is comprised of 17 aerodynamic and physiological descriptors, phonation being one of them. For none of them automatic labelling has ever been attempted; for none of them do annotated datasets exist. This chapter suggests an incremental approach to the complex task of automatic classification of vocal physiology. We begin with a manageable MIR problem setting: a proof-of-concept experiment on automatic extraction of phonation modes from recordings of sustained vowels is presented. Its generalisation is discussed in Section 3.4.

Johan Sundberg in his seminal work "The Science Of The Singing Voice" identifies four different phonation modes in singing: *breathy*, *neutral*, *flow* and *pressed* (Sundberg 1987, chapter 4). Phonation modes characterise the way the sound is produced at the vocal folds, the aerodynamic properties of the vocal source (2.1.1.1). They play an important role in singing: they are an essential characteristic of a singing style; they are utilised as a means of expressive performance; they can be indicative of voice disorders; subtle changes in phonation mode production are assessed routinely by singing teachers to determine the progress of a student (2.1.1.2).

In this chapter we outline our method for supervised classification of phonation modes in Section 3.1. Section 3.2 describes the dataset which we created in order to test this method, by means of the experiment presented and discussed in Section 3.3. Section 3.4 examines a generalisation strategy to a more varied data, that would ultimately lead to the implementation of the revised Cantometrics experiment, and considers its challenges.

3.1 Methodology

Generally in MIR, automatic detection of high-level musical qualities such as phonation modes, keys or genres is achieved in a two-step process. First, low-level audio

features are extracted from music recordings; this step can be thought of as compressing original data into a much smaller sample which still retains the relevant information. Second, a machine learning or other statistical classification method is applied to determine which low-level features correspond to which high-level classes.

To implement this approach using a supervised learning classification algorithm in the statistical component, a so-called *training dataset* is required. It is a collection of audio recordings with semantic labels attached to audio tracks or fragments indicating the high-level classes (such as 'key: D major´ or 'phonation mode: pressed´) to which this audio belongs.

A training dataset was specifically produced for this experiment and is described in detail in Section 3.2. Our feature selection strategy is discussed in the next subsection. For the statistical component we use Support Vector Machines with a 10-fold cross-validation employed for performance evaluation. Parametrisation of the models and automation of the approach are outlined in Section 3.1.2.

3.1.1 Feature extraction

In choosing the low-level feature for our experiment we had to account for the fact that phonation modes result primarily from the glottal activity and are less affected by the form of the vocal tract. In contrast to live singing, where phonation modes can be determined through measurements (using the Rothenberg mask, Rothenberg 1973, or indirectly by means of non-invasive electroglottographs, Howard 2010, Pulakka 2005), for audio recordings of previous events these techniques are not applicable. In this case, either the voice source waveform can be estimated or expert listeners such as phoniatricians and singing teachers can be surveyed to label recording samples with corresponding phonation modes. For an automated solution we have opted for the first approach.

We took Gunnar Fant's source-filter model of sound production as a basis, which assumes that the voice excitation and the vocal tract are linearly separable (Fant 1960). The volume velocity of airflow through the glottis (the space between the vocal folds), the *glottal flow*, is the excitation source for voiced speech and singing. The voice source signal, i.e. the glottal flow, is filtered by the vocal tract to yield the airflow at the mouth; this airflow is then converted to a pressure waveform at the lips and propagated as a sound signal (see the upper row of Figure 3.1.1). The source-filter model assumes that glottal airflow is controlled mostly (though not entirely) by glottal area and subglottal pressure, and not by vocal tract acoustics.

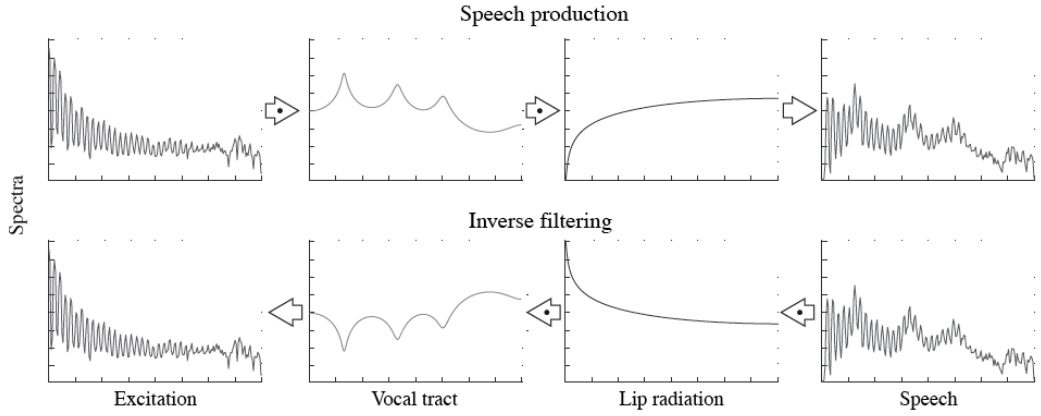


Figure 3.1.1: Inverse filtering: the upper row represents the separated speech production model; the lower row illustrates the corresponding inverse filtering process, in which the lip radiation and vocal tract filters are inverted to acquire an estimate for the glottal flow waveform (reproduced from Airas 2008, p. 50, with permission by Informa Group)

It has been shown that in reality the voice source and the vocal tract interact, and the interaction is even vital in supporting the vocal fold vibration. Thus the source-filter theory should be considered a simplification of the actual voice production process (Rothenberg 1980, Childers and Wong 1994). However, despite its theoretical shortcomings, it is being widely used for speech analysis and re-synthesis in mobile phone transmission, for lossless audio compression such as MPEG-4 and FLAC, as well as in many research studies.

Nevertheless, assuming separability of the model components, an estimate of the glottal flow can be acquired by removing the effects of the estimated vocal tract and the lip radiation from a measured airflow or pressure waveform. This process is called *inverse filtering* (Fritzell 1992, Walker and Murphy 2007, Drugman, Bozkurt and Dutoit 2012, Gudnason, Mark R.P. Thomas and Naylor 2012). The vocal tract (throat, mouth and in some cases nose) forms the tube, which is characterised by its resonances. The resonances of the vocal tract give rise to formants, or enhanced frequency bands in the sound produced. Inverse filtering can be considered roughly as the process of removing the formants (Figure 3.1.1).

A number of publications dedicated to detection of pressed and breathy phonation modes (mostly in speech) employed descriptors derived from the glottal flow waveform such as amplitude quotient (AQ), normalised amplitude quotient (NAQ) and the difference between the first two harmonics (H1-H2) (Walker and Murphy

2007, Orr et al. 2003, Drugman et al. 2008, Lehto et al. 2007, Sundberg et al. 2004). These descriptors are considered particularly suitable for glottal flow waveform estimation because they are relatively robust to some estimation errors. While these coefficients were found useful for phonation mode estimation, there is no explicitly defined correspondence between their values and phonation modes; thus a classification method needs to be employed to detect the implicit relationship.

3.1.2 Parametrisation

Long established implementations of glottal flow waveform estimation require an extensive manual parametrisation with a large number of input values (Granqvist 2003). Fortunately, in recent years semi-automatic and automatic algorithms have been introduced. We opted for a semi-automatic approach called Iterative Adaptive Inverse Filtering (IAIF) (Alku 1992). It requires a manual setting of two input parameters: the number of concatenating segments to model the form of the vocal tract and the lip radiation factor. This algorithm showed a performance comparable to that of a well established manual method (Lehto et al. 2007). A publicly available Matlab package called *TKK Aparat* by Matti Airas (2008) that implements IAIF offered us a platform for further development. We optimised the values of the input parameters via grid search. The optimisation criteria were, in order of importance: classification accuracy; results stability (low standard deviation); and model simplicity.

Interestingly, parametrisation of an IAIF model corresponds to physical properties of vowel articulation. The number of vocal tract segments determines the complexity of the vocal tract form in the model; lip radiation factor is related to lip and mouth opening. Thus, acquiring optimal values for input parameters means parametrising the physical model of articulation. This fact also constitutes a limitation of this modelling approach – for each articulation class a separate model has to be produced.

It is obvious that different vowel sounds require different articulation: while A is wide open, U is quite closed; while for O the mouth is rounded, for I it is flattened. The situation is less obvious for one vowel sound sung at different pitches: though the mouth is usually opened wider at high pitches, the differences in the middle region are usually less significant. Considering utterances of the same vowel in different phonation modes, the variation in articulation depends on the vowel: while for A it will vary only slightly between phonation modes, articulation of sounds like

I and U in pressed phonation differs from that in breathy and neutral phonation considerably. One should therefore expect at best blurred results if different pitches are represented in the same training set.

Our current experiment is based on the assumption that there is only slight variation in articulation for the same vowel sung at various pitches in various phonation modes. Though only an approximation, it has allowed us to make the first step to the solution of the general problem of automatic phonation mode extraction.

3.2 The dataset

For our experiment we constructed a dataset of audio recordings of sustained vowels which is described in this section. While datasets on phonation modes in speech exist, such resources for singing are not available. Our dataset closes this gap and offers researchers in various disciplines a reference and a training set. It is available online under a Creative Commons license through the Open Science Framework repository at `https://osf.io/pa3ha/` [1].

3.2.1 The recordings

The dataset consists of 763 WAV files. Each file contains a single recording of a sustained sung vowel. Recordings are of 1 sec length on average. 500 ms around the middle of the samples were considered suitable for analysis—they displayed a relative stability in pitch, intensity, phonation and articulation (beginnings and ends of the samples are often less stable).

The vowel sounds represented on the recordings are listed in Table 3.1. These sounds were sung on all pitches on a semitone scale from A3 to G5, in every phonation mode given in Table 3.2.

All the recordings were produced by one female singer . This excludes any variation that would necessarily arise between singers, which is useful particularly at the initial stages of classification model training and testing. The singer was professionally trained, with expertise in Western popular and in Russian traditional singing and a profound experience in a number of other music traditions.

The singer's vocal range is approximately D3—C6, with the working range being usually limited to G3—F5. At both extreme ends of the range, phonation became unreliable and the corresponding recordings were not included into the dataset.

[1] Since November 2016 it is being archived, curated and discussed under Open Science Framework.

Table 3.1: The vowels represented in the Phonation Modes Dataset.

Sound (IPA notation)	examples	Symbols used in the labels
[a:]	/a/ – low front unrounded sound, as in English *f**a**ther*, German *R**a**t* or in Russian *м**а**м*	A
[e:]	/e/ – high-mid front unrounded vowel, as in English *g**e**t*, German ***E**sel*, Russian *м**е**сто*	E
[i:]	/i/ – high front unrounded, as in English *fr**ee***, German *Gen**ie***, Russian *в**и**д*	I
[o:]	/o/ – high-mid back rounded, like in German *r**o**t*, Russian *к**о**м*, somewhat similar to English *c**au**ght*	O
[u:]	/u/ – high back rounded, German *F**u**ß*, Russian *л**у**г* , somewhat similar to English *b**oo**t*,	U
[ø:]	High-mid front rounded vowel, as German /ö/ in *sch**ö**n*	OE
[y:]	High front rounded sound, as German or Turkish /ü/, e.g., in German *m**ü**de*	UE
[ɛ:]	Low-mid front unrounded, German /ä/ as in ***Ä**hre*, Russian /э/ like in *э**т**от*, similar to [æ] in English *c**a**t*	AE
[ɨ:]	High central unrounded vowel, Russian /ы/ as in *м**ы***, similar to English *ros**e**s*	Y

Pitches around the singer's register break (D5 to F5) can also be less reliable. We decided to include vocalisation above the break into the dataset to make it more representative, thus all pitches up to G5 were included.

The pitch and the phonation mode were controlled by the singer during the recording. Recordings were examined at a later date and those with the best pitch matches were retained. A subset of the recordings was sent to Prof. Sundberg to control for a correct phonation modes production, and he gave his approval. Unfortunately, as the singer learnt later when attending Sundberg's summer school, she misunderstood the flow phonation mode from reading his book Sundberg (1987). For this reason the flow phonation recordings she produced were not accurate and were discarded in this experiment. They were retained in the dataset though for other possible uses such as vocal synthesis.

Table 3.2: The pitch range of vowels in the Phonation Modes Dataset.

Vowels	breathy	neutral	flow*	pressed
A	A3 - G5	A3 - G5	A3 - H4	A3 - C5
E	A3 - G5	A3 - G5	A3 - H4	A3 - C5
I	A3 - G5	A3 - G5	A3 - H4	A3 - C5
O	A3 - G5	A3 - G5	A3 - H4	A3 - C5
U	A3 - G5	A3 - G5	A3 - H4	A3 - C5
OE	A3 - G5	A3 - G5	A3 - H4	A3 - C5
UE	A3 - G5	A3 - G5	A3 - H4	A3 - H4
AE	A3 - G5	A3 - G5	A3 - H4	A3 - C5
Y	A3 - G5	A3 - G5	A3 - H4	A3 - H4

* Recordings of flow phonation were later found to be inaccurate; they were dropped from the experiment, but kept in the dataset for other uses.

3.2.2 Recording conditions

The recordings were made with a professional dynamic microphone from Electro-Voice, model no. N/D357A. The model was chosen because of its flat response: $+10dB \pm 1dB$ between 200 Hz and 15000 Hz (Figure 3.2.1). The microphone was positioned horizontally at the level of the singer's mouth, at the distance of 100 cm at which the response curve given in Figure 3.2.1 was measured. Svec and Granqvist (2010) give detailed instructions on the choice and positioning of the microphone.

For digitisation of the analogue signal MobilePRE USB was used—a USB bus-powered pre-amplifier and audio interface from M-Audio. It was then connected to a MacBook Pro (early 2008) via USB and the digital signal was recorded using the audio processing software Audacity.

We chose 96 kHz sampling rate and 24 bits precision in compliance with the recommendations for acoustic analysis and archiving by the International Association of Sound- and Audiovisual Archives (IASA TC-04, see Bradley 2009). The recording session took place in a quiet room environment. The requirement of a signal-to-noise ratio of at least 15 dB was adhered to (Svec and Granqvist 2010).

3.2.3 The dataset availability

The dataset is available for download at `https://osf.io/pa3ha/`[2] under Creative Commons CC BY-NC-SA license. This license allows free sharing of the dataset as

[2]Since November 2016 it is being archived, curated and discussed under the Open Science Framework.

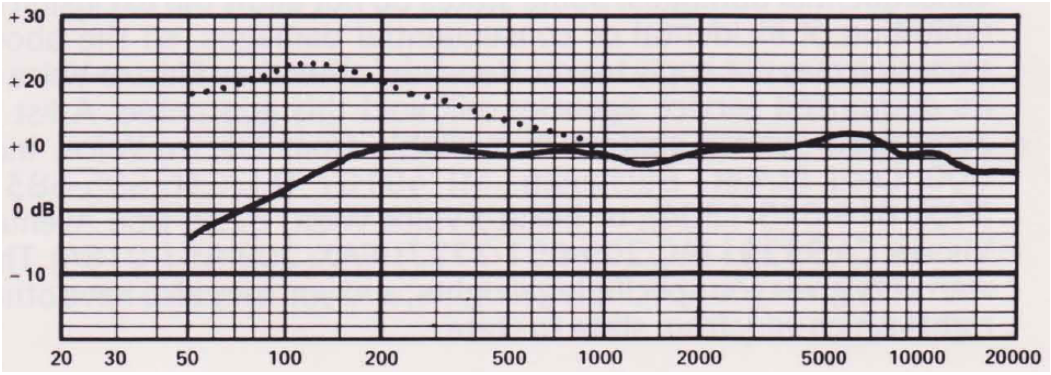


Figure 3.2.1: N/D357A microphone frequency response curve (thick curve). The thin curve marks the proximity effect which only takes effect at the distance of 12 inches (30 cm) or closer.

well as altering it or building new work based upon it. There are following conditions for the use of the dataset according to this license:

- attribution – reference the creators;
- no commercial use;
- share alike – if you alter, transform or build upon it, you may distribute the result only under the same license.

The dataset has been used by several researchers, by some for its original goal – building models for automatic extraction of phonation modes, by others for other aims, e.g. vocal synthesis. Further discussion and related papers by others can be found on the Phonation Modes Dataset project website[3].

3.3 The experiment

The experiment we present investigates the performance of automatic phonation mode classification for nine vowels. For each vowel there is a dataset that contains variation in pitch and in phonation mode only, while other parameters like recording conditions or singer-specific articulation are controlled. The goal of this experiment is to demonstrate that phonation mode detection can be automated for sustained sung vowels and to study the limitations of such an automation. Our methodology is discussed in Section 3.1.

[3] `https://osf.io/pa3ha/`

3.3.1 Experiment design

Because of the model constraints outlined in Section 3.1.2 the experiment was performed separately for each vowel. We decided to use recordings in the pitch range between A3 and C5 only. There is a number of reasons for this: first, the dataset becomes more balanced between phonation modes, because for pitches above C5 only breathy and neutral phonation was recorded; second, the variation in articulation between pitches for a given vowel is minimised; third, including the register break in the training set seems problematic, because the values of the low-level features are likely to change abruptly at the register transition; and fourth, estimating the voice source signal through inverse filtering may become less reliable for higher pitches with a smaller number of harmonics in the spectrum. Thus, for each of the nine vowels we had a training set covering all pitches between A3 and C5 and all phonation modes (with the exception of the pressed mode for vowels 'UE' and 'Y', which are represented at all pitches except C5).

The flow chart of the experiment is given in Figure 3.3.1. First, a grid search for lip radiation and number of vocal tract segments is laid out. For each point of the grid, the voice source waveform is estimated by means of IAIF algorithm with the input arguments given by the chosen point of the grid. Six low-level features are calculated from the estimated waveform. These are then fed into the libSVM implementation of radial basis function kernel SVM. SVM parametrisation is again solved by means of a grid search: first, a grid for C and gamma is laid out; second, for each point of the grid, a 10-fold cross-validation is performed utilising the six low-level features returned by IAIF and the phonation mode labels from the dataset; the mean classification accuracy is returned. The pair of C and gamma producing the highest accuracy value is picked. This best accuracy value is then mapped back to the lip radiation * number of vocal tract segments grid point used for feature extraction. Calculating best accuracy for each combination of lip radiation and number of vocal tract segments in this way constructs an optimisation function in the space spanned by their domains. These optimisation functions were manually studied for each of the vowels to pick a stable maximum and to avoid overfitting.

For feature extraction we used an implementation of the *IAIF* algorithm (see Section 3.1.1) by Matti Airas called *TKK Aparat* (Airas 2008), which is available to download online. We modified the code to allow for batch processing. We enabled the automatic low pass filter, where frequencies lower than f0 are filtered out. We used the samples of 30 ms length for analysis (this parameter is called *selection* in

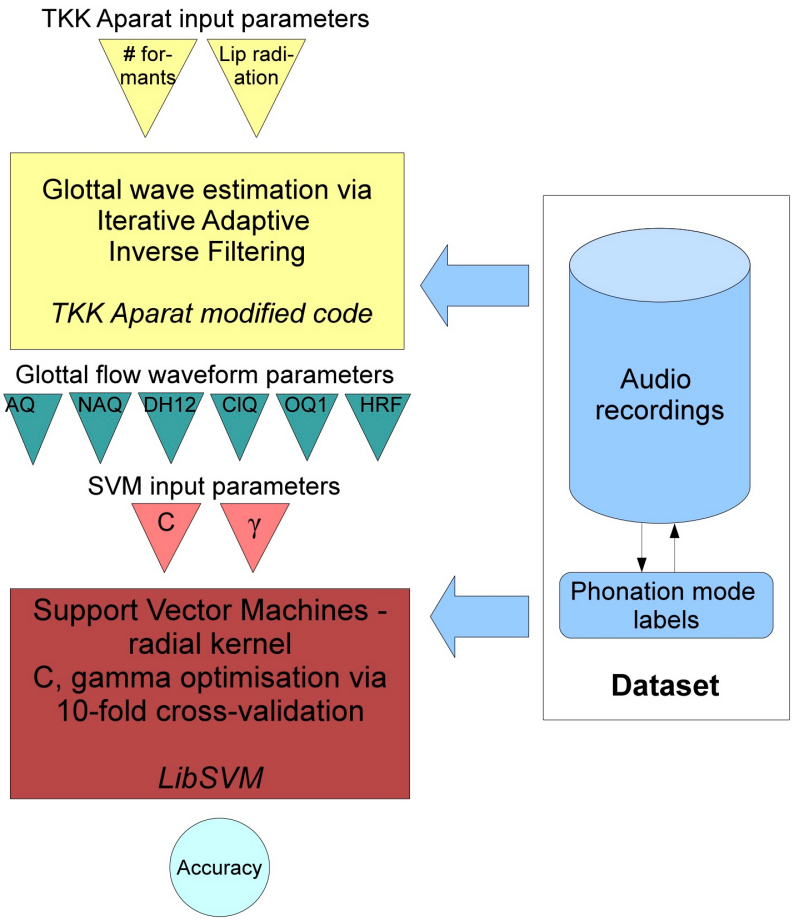


Figure 3.3.1: Experiment flow chart. Our experiment utilises the standard MIR two-stage strategy for automatic extraction of a high-level musical descriptor, consisting of a low-level audio feature extraction and a statistical classification. For the low-level feature extraction we use the IAIF algorithm implemented in TKK Aparat. It requires two input arguments: lip radiation and number of vocal tract segments (the latter denotes the number of concatenated tubes in the vocal tract model). For statistical classification Support Vector Machines with a radial basis function kernel are used, again requiring two input arguments: C and gamma. The values of the input arguments for each of the algorithms are optimised by means of a grid search.

TKK Aparat). This value for the length of the analysis window was determined empirically, as a trade off between the processing time (too long for long samples) and the amount of information contained in the sample. The default value of 20 ms in TKK Aparat was too short, in some instances f0 could not be calculated.

TKK Aparat implementation of the *IAIF* algorithm requires two input arguments, which are called *lip radiation* and *number of formants*. While the term *lip radiation* is applied similarly in the literature on inverse filtering, the use of the term *formant* in *number of formants* by TKK Aparat is misleading: it does not in fact refer to the formants of the vocal tract filtered out by inverse filtering, which is rather determined by the frequency resolution. Instead it denotes the number of concatenated tubes of various diameters used to model the form of the vocal tract. We refer to this parameter as *the number of vocal tract segments*.

The allowed range for the number of vocal tract segments is between 4 and 30. We implemented a grid search between 5 and 29. For lip radiation the range is not limited by TKK Aparat (it only checks that the value is above zero). The default value is 0.99. The values for lip radiation used in speech processing are usually between 0.95 and 1.0. Since the mouth is often opened wider during singing than in speech, our grid search runs between 0.9 and 1.0 with the step 0.005.

TKK Aparat extracts a number of time-related and frequency-related glottal flow waveform descriptors. We use six of them as our low-level features:

1. *Amplitude Quotient* (AQ) is defined as the ratio of the flow peak-to-peak amplitude and the minimum peak of the pulse derivative

2. *Normalised Amplitude Quotient* (NAQ) equals AQ normalised by dividing it by the period length

3. *Closing Quotient* (ClQ) measures the ratio of the duration of the closing phase to the period length

4. *Opening Quotient* ($OQ1$), the time between the primary opening instant and the closing instant normalised by the period length

5. $H1 - H2$ ($DH12$), the difference of the first and second harmonics of the glottal flow waveform spectrum in decibels

6. *Harmonic Richness Factor* (HRF), which is the ratio between the sum of the magnitudes of the harmonics above the fundamental frequency and the

magnitude of the fundamental in decibels:

$$HRF = \frac{\sum_{k \geqq 2} H_k}{H_1}$$

For the details of the glottal flow waveform descriptors see Airas 2008. Figure 3.3.2 shows the distribution of the six voice source waveform descriptors for each phonation mode.

For the statistical component of our experiment we use the *libSVM* implementation for Support Vector Machines in Matlab (Chang and Lin 2001). We employ radial basis function kernel SVM, the values of C and *gamma* are optimised via grid search and passed to *libSVM*. Grid search was implemented in two steps, with a coarse grid search providing an overall picture, followed by a fine grid search around the maxima of the optimisation function on the coarse grid. The optimisation function is the mean classification accuracy of a 10-fold cross-validation. The Matlab code is available on the Phonation Modes Dataset website of the Open Science Framework `https://osf.io/pa3ha/` .

3.3.2 Results

First, a coarse grid search for optimal values of number of vocal tract segments and lip radiation was performed, in order to obtain the shape of the classification accuracy function over the parameter space (Figure 3.3.3). When picking the end result points from several maxima we took in account along with classification accuracy also the stability of the result expressed in standard deviation, and the simplicity of the model, which is reflected in the number of vocal tract segments. For 'I', 'O', 'U', 'Y' the results were blurred, there were one or more areas with high accuracy values in the coarse grid. Here we opted for the more stable results. For 'A' we chose of two maxima a solution which was more stable and had a smaller number of vocal tract segments. At the same time, for 'E' and 'AE' the maxima with the high number of vocal tract segments seem to be genuine and not to result from overfitting, this is supported by a relatively low standard deviation.

Fine grid search results with the corresponding optimal values of input parameters are given in Table 3.3. The average accuracy of over 50% and for all but two vowels of over 60% was achieved, which is well above chance (25% for a four-class classifier).

Table 3.3 also gives the values of the input parameters leading to the highest phonation mode classification accuracy. These optimal values for lip radiation and

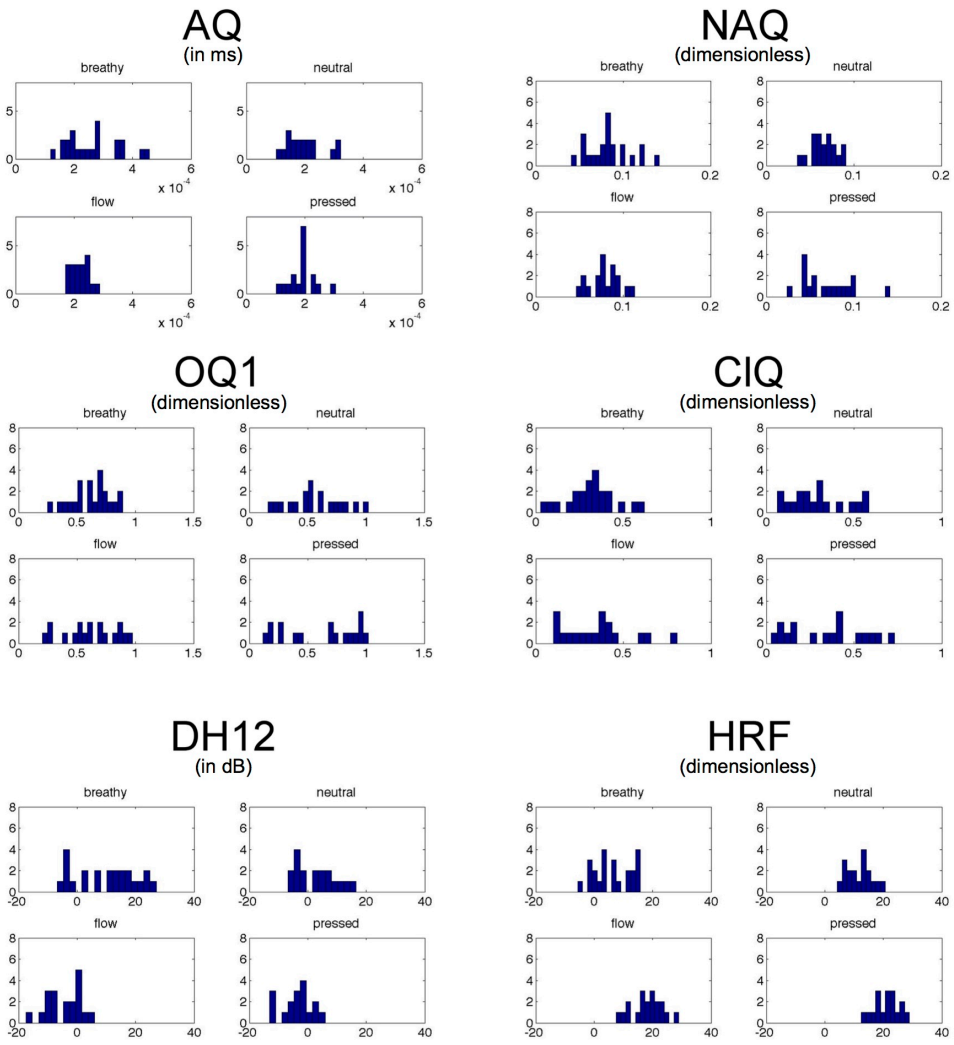


Figure 3.3.2: The distributions (counts of samples) of the six voice source waveform descriptors for each phonation mode for the vowel A.

Table 3.3: Fine grid search results. For each vowel the average accuracy of a four-class classification and its standard deviation in a 10-fold cross-validation are given together with the optimal values of the input parameters: lip radiation and number of vocal segments for IAIF/TKK Aparat and C and gamma for Support Vector Machines/libSMV. Also, the number of files in the corresponding training sets is indicated.

The average accuracy of over 50% and for all but two vowels of over 60% was achieved, which is well above chance (25% for a four-class classifier). These results demonstrate that there is structure in the data.

Vowel	A	E	I	O	U	OE	UE	AE	Y
accuracy in %	61.3	66.4	73.3	67.1	54.8	62.1	65.2	56.9	69.3
std in %	11.6	13.1	12.4	11.8	14.7	14.3	16.8	16.0	12.2
# vocal tract segments	22	28	7	13	23	22	22	29	28
lip radiation	0.91	0.91	0.925	0.91	0.945	0.935	0.935	0.93	0.94
$\log_2 C$	4.25	5	8.75	10	9.75	14	1.75	10	4
$\log_2 \gamma$	-0.25	1	-2.5	-4	-3	-3.75	1.75	-1	-0.5
# training files	77	68	68	70	62	69	67	69	68

number of vocal tract segments – the input parameters of the IAIF algorithm / TKK Aparat implementation – are plotted together in Figure 3.3.4 to allow comparison. Confusion matrices for classification with the optimal input parameters are given in Figure 3.3.5.

3.3.3 Discussion

The results clearly demonstrate that there is structure in the data and that our approach is justifiable. We reached classification accuracy values of 65% on average for a four-class classifier, and standard deviation was in most cases under 15%. At the same time, the structure in the data is blurred for most vowels, which was expected due to assumptions discussed in Section 3.1.2.

Figure 3.3.4 shows that optimal lip radiation values correspond approximately to the relative mouth opening in the production of the vowels: while 'A' and 'E' require a wide open mouth, for producing 'U' and 'OE' the mouth is almost closed. Since this knowledge was not part of the model, it is a further argument that justifies our approach. Thus, as a side effect of our investigation, we have produced (indirect) evidence of physiological properties of the tested vowels – the opening of the mouth and the complexity of the form of the vocal tract.

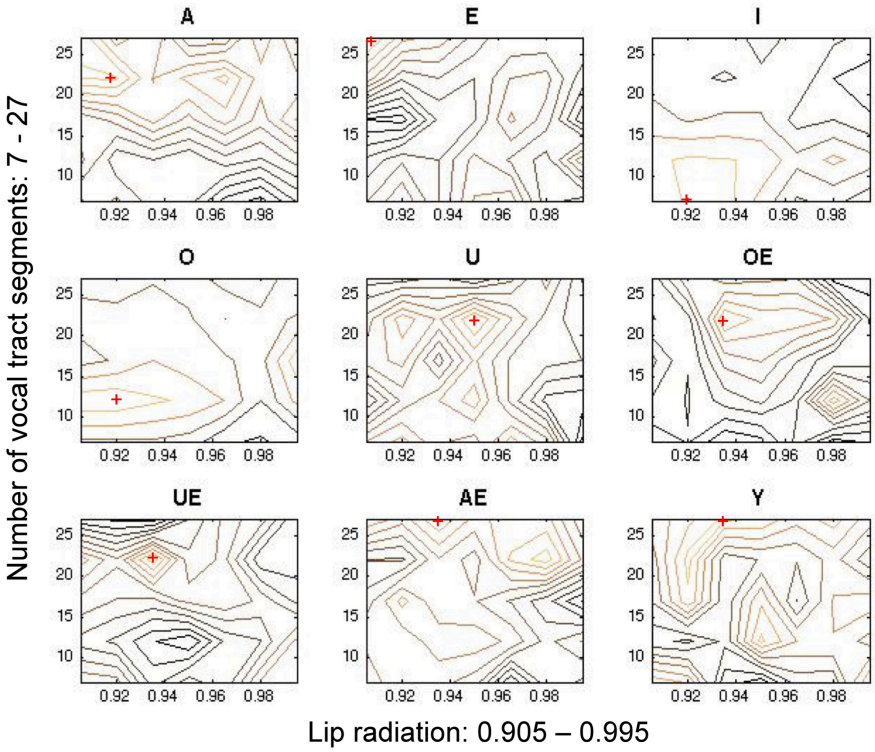


Figure 3.3.3: Coarse grid search results. The graphs represent phonation mode classification accuracy as a function of lip radiation (x axes) and number of vocal tract segments (y axes). Number of vocal tract segments was iterated from 7 to 27 in 5-steps; lip radiation from 0.905 to 0.995 in steps of 0.015. The darker (black) colours represent lower values of the accuracy function, with the maxima in lighter (golden) colours.

A maximum of this accuracy function would represent optimal values of the IAIF input arguments - lip radiation and number of vocal tract segments - for a classification model for a given vowel. For most of the vowels results are blurred, with several maxima or larger areas of high accuracy function values. This was expected due to simplifying assumptions discussed in Section 3.1.2. Optimal solutions (red crosses) were picked manually taking into account besides the accuracy values also the stability of the solution (expressed in standard deviation) as well as the simplicity of the model.

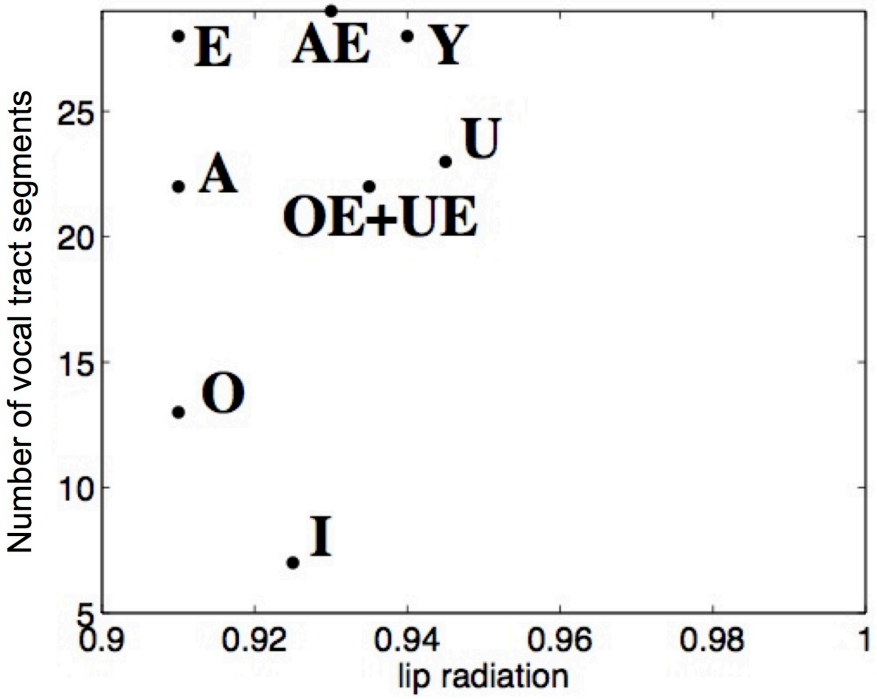


Figure 3.3.4: Optimal solutions for all vowels. For each of the nine vowels, the values of the IAIF input arguments - lip radiation and number of vocal tract segments - leading to the highest phonation mode classification accuracy have been plotted in one space to allow comparison. The optimal lip radiation values found in our experiment roughly correspond to the respective mouth opening during singing of the given vowel: e.g., the lip radiation for A is smaller (mouth opened wider) than for I, which is in turn wider and has a smaller lip radiation than U. The confirmation of these physiological facts by our findings is an indirect justification of the validity of our approach, which did not include any prior physiological knowledge.

real \ predicted	Breathy	Neutral	Flow	Pressed
A				
Breathy	14	7	2	0
Neutral	6	11	1	0
Flow	3	1	10	5
Pressed	0	1	4	12
E				
Breathy	15	3	1	0
Neutral	4	9	3	1
Flow	1	2	9	4
Pressed	0	1	3	12
I				
Breathy	14	4	0	1
Neutral	3	13	1	0
Flow	0	2	10	3
Pressed	0	0	4	13
O				
Breathy	16	2	1	0
Neutral	2	13	2	0
Flow	1	1	12	4
Pressed	0	1	9	6
U				
Breathy	15	2	1	1
Neutral	8	7	1	1
Flow	0	1	7	5
Pressed	1	2	5	5
OE				
Breathy	16	2	1	0
Neutral	4	10	1	2
Flow	2	2	10	3
Pressed	0	5	4	7
UE				
Breathy	13	5	0	1
Neutral	3	12	2	0
Flow	1	3	8	4
Pressed	0	0	4	11
AE				
Breathy	15	3	1	0
Neutral	3	9	3	3
Flow	1	2	9	4
Pressed	2	2	6	6
Y				
Breathy	19	1	0	0
Neutral	5	10	2	1
Flow	2	2	9	2
Pressed	0	2	4	9

Figure 3.3.5: Confusion matrices for phonation mode classification. There is more confusion within two subgroups: breathy+neutral and flow+pressed, than there is between these subgroups. Breathy phonation can be clearly distinguished from pressed in the vast majority of cases.

Interestingly, the vowels 'OE' and 'UE' display the same optimal lip radiation and number of vocal tract segments. This means that one IAIF model can be used for phonation mode detection in samples containing both vowels. The optimal input values for other vowels differ, thus different models have to be used for each of them.

Confusion matrices demonstrate that there is generally more confusion between breathy and neutral modes than there is between breathy and pressed or between neutral and pressed. For a better understanding of the model limitations, a detailed misclassification analysis would be instructive.

To improve results of the presented experiment, the quality or the quantity of the data would probably have to be extended. If the recordings are made with a special condenser microphone suitable for acoustic research (see Svec and Granqvist 2010) and if the sound pressure level is measured and documented during the recording event (see, e.g., Fritzell 1992), higher quality glottal flow waveform estimations can be achieved. A more diverse training set, on the other hand, would result in more robust classification. Also, if enough recordings are available for each vowel and each pitch, differentiation of the IAIF model in respect to pitch (such as wider mouth opening at higher pitches) can be taken into account and investigated.

The obvious limitation of the chosen approach is the dependence of the IAIF model on the physiological properties of the vowels. This implies that if phonation mode detection is attempted on real-life recordings, a component extracting and detecting vowels has to precede feature extraction. This may introduce additional errors and have a negative impact on the overall result. Alternatively, automatic inverse filtering approaches can be applied, though their performance might be inferior to IAIF. Further generalisation suggestions are given in Section 3.4.

The issue of acquiring more reliable ground truth will have to be considered. Currently the labels in the dataset are based on the singer's understanding of what phonation mode was sung. Ideally, in future, ratings from a larger number of experts should be obtained. Apart from that, a new set of recordings could be produced with EGG measurements gathered during recording. These measurements would provide an additional argument in determining the phonation mode of a singing sample. A more objective verification of phonation mode labels for one dataset would build up a golden standard, allowing to test future experts who will rate new datasets and thus expand the whole area of research on phonation modes.

In the presented experiment we have employed a method for automatic extraction of phonation modes from idealised data. The problem of application of filtering, source separation and other techniques from the signal processing literature to adjust

the application of our method to real-world data is reserved to future work.

An interesting subject for a future investigation, that was touched upon in our work, is the relationship between phonation modes and registers. Though a whole chapter of Sundberg's book is devoted to registers, he does not specifically discuss the register in regard to phonation modes. He neither states explicitly that his definitions of phonation modes are only operational for the modal register or just the chest voice, nor does he mention how they would work in the falsetto register or at the register break. In our practice of singing performance and teaching, flow and pressed phonation are difficult to produce in the range close under the register break. In Western classical singing school flow phonation is used extensively in the chest voice and in the head voice, while at the register break a technique called *covering* is used based on neutral phonation to mask the transition between registers. In other repertoires such as musical theatre *belting* is used instead, where the closed quotient seems to be longer, which points to more pressed phonation. Though the mechanism of change in the perceptual quality of the voice at the chest voice/head voice transition is still debated, this part of the range is sometimes denoted as *mix*, *voix mixte*, etc. (see Section 2.1.2). Could that involve mixing phonation modes as well?

3.3.4 Confounding issues and further work

Our dataset has been used by others for further research. In particular, two groups of researchers approached phonation mode classification with recent MIR techniques.

Rouas and Ioannidis (2016) find that acoustic features perform better than glottal features in separating phonation modes. Stoller and Dixon (2016) investigate a variety of MIR features in relation to their classification performance for the task. They confirm the former finding, demonstrating, among other things, that MFCCs can better classify phonation modes than our glottal features. Both papers report similar results achieving about 80% classification accuracy.

Because phonation modes are defined by Sundberg as a function of vocal source and not of the audio signal propagated from the mouth, the success of conventional MIR methods in comparison to our vocal source based model raises a question about possible confounding issues. These may include:

a) a low precision of the voice source waveform de-convolution. This is certainly an issue, in particular with automatic approaches to inverse filtering used in the papers.

b) non-linear aerodynamic effects in the process of vocal production. It has been an area of increased interest among voice and singing researchers. It has been shown that non-linear effects play an important role (Titze 2008, Butte et al. 2009).

c) limitations of a one-subject dataset

d) over-fitting, classifying a different feature. In particular, Stoller and Dixon mention loudness/overall energy as one important factor that differentiates the recordings in our dataset.

Cepstral coefficients are widely used in MIR but have a disadvantage of being non-transparent and hard to interpret. Stoller and Dixon present a visual analysis of MFCCs in their task, showing that higher coefficients are dependent on pitch. It could be assumed that lower coefficients capture information about vocal source including the non-linear effect of vocal tract on phonation.

Rouas and Ioannidis recorded their own dataset similar to ours, sung by a professional baritone singer. Their classification results were similar for each of the datasets and for a combined dataset, thus addressing the one-singer limitations and demonstrating that their method is not overfitting. They also used their classification for further high-level analysis, attempting style differentiation for the same singer, as well as detection of laments in a dataset of singing examples from several cultures.

3.4 Revising Cantometrics – incremental approach

We pointed out in Section 1.4 of the Introduction that in setting ourselves the task of revising Cantometrics we are facing one of the most general MIR undertakings: automatic extraction of high-level descriptors from a highly heterogeneous audio dataset. In this Chapter we consider an incremental approach to this task. We started with controlled conditions, where variation was introduced in pitch and in phonation only. There was just one singer, thus no inter-singer variation was present; recording conditions were the same for all tracks; the singer sang only sustained vowels, so non-harmonic sounds like consonants did not interfere. In the previous sections of this Chapter we described a proof-of-concept experiment under theses conditions that demonstrated the presence of structure in our data.

Further research by others confirms that this task is solvable on MIR terms (Section 3.3.4). Yet it is the generalisation to more complex tasks that will pose the biggest challenge. For each generalisation step a new training set (annotated dataset) will have to be created covering a new source of variability in the data. The

generalisation steps for phonation mode extraction could be outlined as follows:

1. Include multiple singers; diversify by gender, age, anatomy, vocal proficiency, musical tradition/genre. This step will counter possible overfitting issues in the current experiment.

2. Move from sustained vowels to real-life vocal utterances (songs). Vocal frames extraction and vowel determination will have to be integrated at this stage prior to phonation modes classification.

3. Move from monophonic solo singing to several singers vocalising simultaneously: duets, trios, ensembles, choirs, monophonic as well as polyphonic

4. Include instrumental accompaniment (voice separation, reliable frames), diversify instruments, musical traditions

5. Introduce varying recording formats and hardware, recording quality.

6. Move away from lab conditions and embrace a variety of recording conditions including the possibly most general setting – ethnomusicological field recordings.

Rouas and Ioannidis (2016) have addressed points 1 and 2 to some extent, adding recordings by another singer to their training set. They have also attempted to use their phonation mode classification on a small dataset of ethnomusicological recordings, differentiating laments from singing.

For each generalisation step a new model will have to be developed, trained and tested based on the new training set. Phonation modes accuracy in the training set should be controlled by independent measurements where possible. At stages 1-5 electroglottographic analysis during recording could provide objective evidence. Yet at later stages expert judgement inevitably will have to be relied upon, and, as we show in Chapters 5 to 7, it is problematic, in particular for the phonation modes descriptor.

It seems to be reasonable to assume that the performance may decrease with increasing complexity. While 80% accuracy was reported for our dataset of sustained vowels (Rouas and Ioannidis 2016, Stoller and Dixon 2016), which is a good MIR result, there are still 20% misclassified recordings, and this amount will likely increase with each step.

Our ontology is essentially a product of 11 descriptors and to classify a recording in accordance to the ontology means to determine the value of each dimension.

Classification errors for each dimension will multiply for the overall result, at which point the accuracy of automatic classification may deteriorate. The simplification path is easy at the start but the difficulty rises with each generalisation step. Future technological developments and algorithmic findings may help overcome the challenges.

4 The mixed-method study – methodology

Motivated by the interest in cross-cultural comparative experiment Cantometrics and its findings about correlations between vocal production and societal traits (Section 1.4), we sought to reformulate Cantometrics subjective vocal production descriptors in more objective terms and to revise and widen its methodology. We developed an ontology of vocal production based on voice source aerodynamics and vocal tract physiology (Chapter 2). Chapter 3 investigated an incremental approach to automatic labelling of our ontological descriptors. It presented a proof-of-concept experiment on phonation modes classification for sustained vowels sung by one singer; outlined the steps for generalising it to more complex data; discussed the challenges related to the generalisation process.

This chapter lays out a methodology for our second – integrated – approach which addresses the whole data variability in the original Cantometrics dataset. We aim to produce reliable annotations of vocal production for that dataset which can then be used for training and testing computational models for automatic rating of our descriptors.

In this chapter we describe a method to collect such annotations (ontology descriptors ratings) in absence of direct measurements. Social sciences mixed-method research techniques for expert knowledge elicitation are employed. We use qualitative and quantitative data collected in interviews with experts to

a) verify the applicability, relevance and completeness of our ontology and adjust it if necessary,

b) examine the consistency of experts' ratings, their agreement about the meanings/values of the descriptors,

c) collect reliable ratings,

d) explore confounding issues and underlying reasons in cases of disagreement.

If we are successful in demonstrating consensus among expert raters, expert knowledge can be claimed a consistent source of annotations and can be used to annotate a larger training set for a general model; such a model will be capable of automatically rating our descriptors for previously unseen recordings of singing, including ethnomusicological field recordings. This result would bring us much closer to the goal of revising Cantometrics and scaling it up to include all music recordings.

In the next sections we explain our study design and follow its iterative implementation.

4.1 The study

To produce a sound the singer inhales air, activates his abdominal, laryngeal and other muscles, adjusts the form of the vocal tract to achieve the sound wanted, thereby exercising an act of a very subtle muscle control and coordination. The difficulty with practical observation arises from the fact that the mechanisms used for this control are hidden within our bodies and can only in part be measured by the instruments currently available. The most common of these instruments include:

- audio recording for the sound waves
- video recording or direct observation of visible body parts
- laryngoscopy, oral and nasal stroboscopy to observe the larynx (and epilarynx)
- X-Ray, MRI to capture position and movement of inner body structures
- electroglottography to measure vocal folds adduction and abduction

etc. (see Kayes 2013, Chapter 2).

The majority of these observation methods have only appeared recently and thanks to them we have learnt a lot about the process of vocal production. Still there are limitations, for example, in MRI picture resolution is low for the temporal scale at which changes in vocal tract happen. Video laryngoscopy allows for much higher resolutions, but it requires a rigid or a flexible scope to be inserted into either mouth or nose, which affects the sound production. While these clinical observation methods have been instrumental in advancing our knowledge of vocal production, they are only applicable during the process of phonation. Recordings of singing (like

in the Cantometrics dataset) cannot be investigated with such techniques. Therefore, while most of our descriptors may be observed with clinical instruments in a lab during vocalisation, which helped to study their mechanism and function, estimating physiological settings from audio recordings of previous vocalisations is an unsolved problem. No direct measurements of our descriptors can be provided for the recordings in the Cantometrics dataset.

We decided to take an investigative practical approach – to interview professionals who use their knowledge of vocal production and of vocal physiology on the daily basis. These experts from different fields and disciplines (singing teachers, medics, speech therapists, voice scientists) are dependent on accurate reconstruction of the physiological process of sound production and have years of experience in this task. We aim to elicit their tacit knowledge by means of interviews and documented analysis of suggested samples of singing. In this study we will be assessing whether our participants can agree about the meaning and the values of our ontological descriptors.

Due to its interdisciplinary nature our study does not fit into the main traditions of enquiry in social sciences; it is closest to knowledge elicitation studies for expert systems and artificial intelligence (Cooke 1999, Ford and Sterman 1998), borrowing elements from phenomenological studies and grounded theory in social sciences (see Creswell 2006). The quantitative component will include collecting ratings for the values of ontology descriptors for a set of audio recordings of singing. The act of analysing the singing samples and rating the descriptors will be used to prompt participants to offer their views and concerns about the descriptors and the ontology as a whole and to suggest improvements. Further questions will be asked to elicit alternatives, salience and relationships between terms and concepts. The semi-structured interviews will guide the participants through the pre-planned rating procedure, leaving space and time for open-ended questions and exploration; in the latter part the focus will be on participants' views with attention to detail and special cases arising. Our questions may change during the process of research to reflect an increased understanding of the problem (see Creswell 2006). While the quantitative analysis will focus on inter-participant agreement, the qualitative analysis will include open coding, an iterative element, presentation of multiple realities.

In more general terms our study is a proof-of-concept investigation into the validity of the physiological approach to modelling vocal production in audio recordings of singing. We are not aware of any research into the accuracy of detecting

physiological configurations through listening. If we assume that our experts are able to detect these configurations reliably, we would expect their ratings to be in good agreement. If a significant tendency to agreement can be demonstrated through quantitative analysis, we can argue that there is in fact an inter-subjective consensus on the meaning of the terms among our experts. Diversification of our participants in terms of profession, research tradition and cultural background would provide a good basis for generalisation, though further studies would be needed to investigate whether other factors would affect the inter-subjective consensus.

In case of agreement the further course of the study would involve dimensionality reduction and investigating possible correlations between our descriptors and Cantometrics parameters. The singing recordings with their ratings would constitute the first annotated dataset of vocal production which can be used as ground truth for computational models that would automatically rate our ontological descriptors for a large variety of singing recordings.

If no agreement is found in the quantitative phase, our study would then focus on qualitative analysis to investigate the underlying reasons for disagreement. Absence of inter-participant agreement would raise questions about experts' ability to determine physiological configurations auditorily-perceptually and therefore the viability of physiological approach to modelling vocal production in general. Claiming that our experts cannot detect physiology through listening would be a serious professional accusation – a detailed investigation of the conditions of the study and possible confounding issues should precede any such claims. Yet absence of agreement should be a warning sign to many lines of research related to singing, meaning that at the current state of knowledge there exist multiple perceived realities in relation to vocal production mechanism and we've got no instruments to determine whether any of them correspond to physical reality.

Given the fuzzy nature of our knowledge and perception of singing we expect to see mixed, fuzzy results at this point. Therefore the study will evolve in an iterative manner depending on what we learn at each stage and an exact planning is not possible.

4.2 Planning the data ahead – number of participants vs. number of musical examples

While designing and setting up our study, we were subject to several constraints:

- we had to collect enough quantitative data to be able to perform statistical analysis – too small a dataset would make our analysis results meaningless
- our qualitative data should not exceed an amount that can be analysed by one person within several months
- we had to find experts in the field of vocal physiology who would be happy to take part in our study
- these experts would have a very limited time they can be asked to contribute to the study
- the group of experts we interview should be diversified in terms of their profession and, if possible, in terms of their cultural and musical background
- the set of musical examples should be diversified in terms of cultures they originate from
- the set of musical examples should contain a large variety of vocal production, covering many points in the space defined by our vocal production ontology

These constraints lead to boundaries for the number of participants and the number of musical examples to be employed for the study. The number of participants should be large enough (particularly if they only have limited time to devote to the study) to produce enough quantitative data; at the same time, the entire interview time should be limited to enable qualitative analysis within our capabilities. Each musical example should be annotated by several participants, which limits the number of examples; yet the examples have to represent the ontological space sufficiently.

Taking all the above into account we decided that the number of participants should be between 10 and 20 and ask the experts for 90 minute interviews – that would leave space for diversification and limit the amount of qualitative data. The next question we had to answer before starting to set up the study was how many musical examples did we have to prepare. Let $Part$ be the number of participants and Tr be the number of musical examples (tracks) to be annotated. Ideally each musical example should be annotated by each participant. This would simplify the statistics. At the same time, the more examples are annotated the better is the coverage of the ontological space. Modern statistics can deal with missing data, therefore we do not necessarily need all experts to analyse every example. We introduced a constraints constant C that gives the minimum number of times we

expect each track to be annotated. If $C = Part$ then each track is annotated by each participant. If C is very low and some tracks are only annotated once or do not get annotated at all, inter-participant agreement cannot be analysed. For our purposes we set $C = 5$, so that for each track inter-participant agreement can be computed and is meaningful, and at the same time more tracks can be presented to participants. Let X_i be a variable showing how many times a track Tr_i is annotated. We'd like the probability of $X_i < C$ to be very small, e.g. lower than 5%:

$$Pr(X_i < C) < 0.05$$

For $C = 5$ we have:

$$Pr(X_i < 5) = \sum_{j=1}^{4} Pr(X_i = j)$$

We do not know in advance how many musical examples participants will be able to annotate within the 90 minutes of an interview. As an estimation we have taken our own experience of annotating musical examples with the dimensions of our ontology, taking into account that experts are more used to this kind of analysis than we are, and at the same time that they are not familiar with the musical traditions from which the examples will be taken. We set the expected number of tracks to be annotated by a participant during an interview to be 7:

$$TrInt = 7$$

The tracks are presented to participants in a random order. The probability for a track to be presented to and analysed by a participant is

$$p = \frac{TrInt}{Tr}$$

Participants perform their analysis of musical examples independently from each other and the order in which tracks are presented to them is independent from other interviews. Whether a track is presented (and analysed) during an interview can therefore be viewed as throwing a coin with the p chance of success. The "coin" is thrown for each participant, altogether $Part$ times. We employ the binomial distribution formula and get

$$Pr(X_i = j) = \binom{Part}{j} p^j (1-p)^{Part-j}$$

Table 4.1: Number of participants vs number of musical examples. This table gives the values of the probability that a track is annotated by less than 5 experts in a study. It is desirable to keep this probability low when planning a study, e.g. below 5%, so that tracks are annotated sufficiently frequently for inter-participant agreement to be meaningful. At the same time a larger number of tracks allows for a better representation of the ontological space in the study. And the lower the number of participants, the more manageable will qualitative analysis be. The optimal values (the highest values below 5%) are marked green. This table allows a preliminary estimation of the tradeoff between the number of participants and the number of tracks. In our study we expected 12-14 experts to take part, therefore 11-12 musical examples was an estimation for a good size of the dataset.

		number of participants										
		10	11	12	13	14	15	16	17	18	19	20
no. of musical examples	10	0.047	0.022	0.009	0.004	0.002	0.001	0.000	0.000	0.000	0.000	0.000
	11	0.112	0.061	0.032	0.017	0.008	0.004	0.002	0.001	0.000	0.000	0.000
	12	0.195	0.121	0.073	0.042	0.024	0.013	0.007	0.004	0.002	0.001	0.001
	13	0.286	0.195	0.128	0.082	0.051	0.031	0.019	0.011	0.006	0.004	0.002
	14	0.376	0.274	0.194	0.133	0.090	0.059	0.038	0.025	0.015	0.010	0.006
	15	0.459	0.353	0.264	0.193	0.137	0.096	0.066	0.045	0.030	0.020	0.013
	16	0.534	0.429	0.335	0.256	0.192	0.141	0.102	0.073	0.051	0.036	0.025
	17	0.598	0.497	0.403	0.320	0.249	0.191	0.144	0.107	0.078	0.057	0.041
	18	0.653	0.559	0.467	0.383	0.308	0.243	0.190	0.146	0.111	0.083	0.062
	19	0.700	0.613	0.525	0.442	0.365	0.297	0.238	0.189	0.148	0.115	0.088
	20	0.738	0.660	0.578	0.497	0.420	0.350	0.288	0.234	0.188	0.150	0.118

The lower bound probability estimation is then given as

$$Pr(X_i < 5) = \sum_{j=1}^{4} \binom{Part}{j} (\frac{7}{Tr})^j (1 - \frac{7}{Tr})^{Part-j}$$

and should not exceed 5%. We have done the calculation for various values of $Part$ and Tr which are summarised in Table 4.1 .

As we approached experts for participation, we created a pool of 15 professionals who were happy to take part in the study. We expected that some interviews may not take place for various reasons. According to the above calculations, 11-12 musical examples were a sensible dataset size for a study with 12-14 participants.

4.3 Musical examples and music analysis

This section outlines our choice of eleven tracks from the Cantometrics dataset (see Lomax 1977). Nineteen physiologically stable fragments were extracted from the tracks, which were then used as entities of analysis in the interviews.

Lomax was very diligent in choosing recordings for his Cantometrics dataset; being one of world's greatest collectors himself, who traveled across all continents, when it came to cultures in which he was not an expert, he would go to considerable lengths to involve leading specialists emic to the culture. We had a chance to convince ourselves of that diligence while studying the Cantometrics' Russian sample: Lomax traveled to Soviet Russia and met with Anna Rudneva, a prominent researcher of Russian musical traditions, and her colleagues. He interviewed Russian musicologists, asked them to choose musical material for him, listened to recommended recordings and to live performances.[1] Certainly, his questions were informed by his hypothesis about the relationship between singing and the society. Also, his approach to compiling the dataset had been criticised by ethnomusicologists, particularly the decision to limit the number of samples representing a culture to 10. Nonetheless, our impression of Lomax's practice of data collection for his experiment was that it was as sound and as unbiased as was possible under the circumstances, and was enhanced by Lomax's true interest in people, in their culture and in their songs.

We took the chapter on *vocal width/vocal tension* from the Cantometrics Training Tapes as our starting point to construct the dataset for physiological analysis (Lomax 1977). The tracks in this chapter were used to illustrate different states of *vocal width*; this descriptor refers to several aspects of vocal production (see section 2.3), therefore, such a choice guaranteed a large variance in vocal quality and a dataset that would cover a wide variety of points in the space of physiological states related to vocal production. All the tracks originated from different cultures thus providing a wide geographic spread. Also, as we mentioned above, Lomax's diligence in choosing tracks for his dataset stood for a solid ethnomusicological approach (See Figure 4.3.2).

Using the Cantometrics Training Tapes as a pool we considered two criteria in choosing musical examples for our study. Firstly, we were looking for fragments that were physiologically stable, with as little physiological changes as possible;

[1] Documented through audio recordings of interviews and sets of chosen samples, can be accessed through Association for Cultural Equity website `http://www.culturalequity.org/lomaxgeo/` (click on Moscow, last accessed 08/09/2017)

our participants were supposed to analyse musical fragments by listening, therefore these stable fragments had to have a length of at least 2-3 seconds to enable auditory perception of complex phenomena. Secondly, we were interested in singing tracks which displayed a single predominant physiological configuration: this would allow us to relate this configuration and participants' ratings of it to the perceptual ratings collected in the Cantometrics experiment.

From 16 musical examples presented in this book chapter we excluded four examples illustrating yodel: in yodel changes in vocal production are deliberate and frequent, making it hard to extract physiologically stable fragments of any significant length.

We further analysed the remaining 12 tracks to determine what we thought were significant physiological changes as well as regions of physiological stability between such changes (see Figure 4.3.1). We were very aware of our own biases and did not intend to impose neither this structure nor our physiological judgements on the participants. Our goal was to prepare entities of analysis that could be associated with a single set of physiological ratings. Also, for the sake of comparability with Cantometrics we concentrated on the tracks that in our opinion displayed one predominant physiological configuration.

While most of the tracks clearly displayed a predominant configuration, which was employed for at least 60% of the track duration, in one track the singer deliberately varied his vocal production and none of his physiological configurations covered more than 30% of the track duration. This track was excluded from our dataset, reducing the number of tracks to 11. A presence of a predominant physiological configuration in a Cantometrics track would allow for a straightforward generalisation of this predominant state to be representative for the whole track, making our participant's ratings results comparable with the Cantometrics ratings.

For the remaining 11 tracks all physiological configurations representing a significant (over 10%) proportion of the track duration were analysed. While three tracks were dominated by one significant configuration, eight other tracks displayed two distinctly different configurations. For each of these 19 physiological configurations representative fragments of the tracks were extracted, which we called *snippets*. Thus each track was represented by one or two snippets. The lengths of the snippets varied from 2.5 seconds to about a minute: for track 30 the snippet contains the whole track. The goal of our analysis was to extract fragments that would allow consistent physiological evaluation.

In Figure 4.3.2 each coloured block represents a Cantometrics track: its number

is given in the first column and its geographic and ethnographic descriptions in the last two columns. The first line in each block gives the duration of the track (column 2) and the duration of singing in the track (column 3). The lines below list the snippets - the representative fragments of stable physiological configurations, extracted from this track for further physiological analysis. Column 2 gives the physical length of the chosen snippet. Snippets are often shorter than the whole duration of the given physiological state, which is given in column 3. Column 4 displays the proportion of the physiological configuration in the whole track. Note track 22 where two snippets represent two parts performed simultaneously, not two consecutive fragments. Blocks are coloured according to their Cantometrics ratings.

It must be noted that salience of particular physiological states is in itself an unstable measure. Tracks in the Cantometrics dataset are usually between 20 and 90 seconds long and are themselves fragments of longer recordings. Lomax was convinced that the main characteristics of singing specific to a given culture are present – are in fact dominant – in any instance of it. He chose fragments of recordings he considered representative of the whole recording. He did not to our knowledge perform a detailed analysis similar to the one described above. In that sense, his choice of fragments was arbitrary, he did not have any measurable criteria for choosing any particular fragment; had he chosen a different fragment, our measures of salience might have been different.

The extracted 19 snippets were used in the interviews as the main entities for physiological analysis. It was important to us to minimise the influence of our choice of snippets on participants' ratings. Therefore, the participants began the analysis by listening to the whole track first; they were asked questions about their perception of singing in the track to allow them to focus on what might be salient to them; only after that snippets were introduced and analysed. Participants were prompted to challenge the choice of the snippets, which in fact happened for track 32. If they did not challenge our choice, we explicitly asked them whether the snippets were representative for the track's physiological states. We were very aware that our judgements of physiological stability were biased – this painstaking procedure of participants' control over our choice of snippets was implemented to counteract that bias.

In most cases the participants were happy to describe each snippet with a single set of ratings. There was one special case though: in track 22 the choir sings in two parts; some participants produced a set of ratings for each of the parts, others stated explicitly that they only rated the upper part, while the rest did not find any

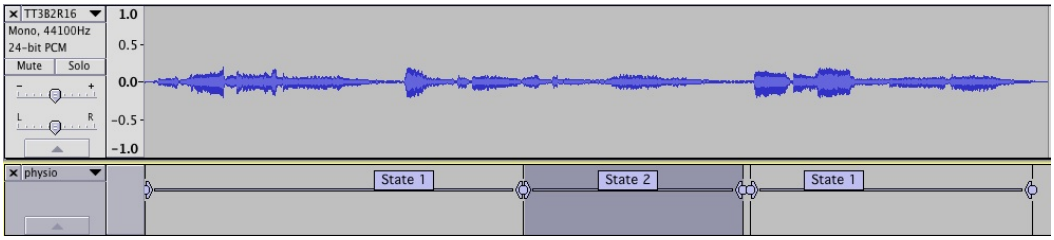


Figure 4.3.1: Analysing physiological configurations in Cantometrics *vocal width* examples. Analysis was performed manually by us and physiologically stable regions were labeled using Audacity software. Track TT3B2R16 from the Cantometrics Training Tapes displays two different physiological states. The first state spans from the beginning to 16 seconds and from 26 seconds to the end of the track. The second state covers the rest - from 16 to 26 seconds. Given the length of the track being 38 seconds, most of which is singing, an approximate proportion of the first physiological configuration constitutes 70% and of the second physiological configuration 30% of the track duration.

physiological differences in vocal production between the parts.

We received the Cantometrics Training Tapes recordings from Alan Lomax Archive in NYC in form of mp3 tracks that were produced during digitisation of audio cassettes, the original collection format. mp3 is a problematic choice of format for acoustic analysis, and at least one participant criticised it explicitly. Unfortunately, it was the only format in which the Cantometrics recordings were available, and we decided that the advantages of having an ethnomusicologically annotated dataset, varied culturally and in terms of vocal production, with attached perceptual ratings and of considerable scholarly interest outweigh the disadvantages. Nevertheless, the question remains open whether physiological ratings would differ for a different recording format.

Audio files containing our musical examples can be accessed at `https://osf.io/6xzg8/`.

4.4 Ontological descriptors

In Chapter 2 we put together an ontology of vocal production based on vocal source aerodynamics, laryngeal and vocal tract physiology (Table 2.2). The ontology includes Johan Sundberg's vocal source parameters, the classical registration terminology and Jo Estill's physiological building blocks of vocal production. It consists

Track/ Snippet	Seconds	Length of singing (sec)	Proportion in the sample	Culture	Lomax's description
16	38.5	32		Galicia, Spain	The sexual mores of Pyrennean Spain, especially Galicia, are notably more relaxed and permissive than further to the South. Here a Muleteer sends his voice through the hills as he travels, calling out for custom and (as he told me) love. P (Lomax #23, B 9)
16_1	16-26	9	28%		
16_2	25.5-38.5	23	72%		
22	31			E. Europe, Slovakia, The High Tatras.	Well-blended choralizing based in liquid, wide-voiced vocal style is relatively common in Eastern Europe, where sexual mores are comparatively permissive. P (Czechoslovakia)
22_1	5 – 14		100%		
22_2	5 – 14		100%		
26	33.5			Maori, New Zealand	The New Zealand Maori, in the general fashion of the Polynesians, preserved their genealogies (so important to them in determining their social positions) in lengthy chanted poems, like this lament for the great chiefs of the past. P (Maori #1, A 1)
26_1	0-9		100%		
32	49	41		Ukraine	A Cossack ballad from the Ukraine, performed by a trained baritone with bandura (stringed instrument) accompaniment. P #1)
32_1	4 – 11.5	16	39%		
32_2	11.5 – 16	21	51%		
14	19	16		Nova Scotia, Canada	A Nova Scotian remembers one of the comic British ballads her ancestors brought with them into the New World. R (Creighton, B 23)
14_1	5 – 10	12	75%		
14_2	9.5 – 14	4	25%		
20	20			Corsica	A Corsican woman sings a lullaby from the not-so-distant time when a bandit, pursued by the gendarmes, was concealed by his wife in her child's cradle. R (Marcel-Dubois and Andral, A 13)
20_1	0 – 9.5		100%		
28	30	22		Arkansas, US	The burry-voiced, off-hand delivery, typical of the Ozarks (Woody Guthrie had it), suits this Rackansack (slang for Arkansas) variant of the English cumulative swapping song. Male solo with mouthbow. P. (Lomax #17)
28_1	1.5 – 9.5	16	73%		
28_1	1.5 – 9.5	6	27%		
34	50	45		Ireland	An Irish folk singer with plenty of "bias" (ornament) performs the most popular ballad in English, the story of how Willy dies for love of a hard hearted girl - (Barbara Allen, here called Mary Ellen) R (Kennedy (& Lomax) #2, B 7e)
34_1	7 – 15	35.5	79%		
34_2	40 – 44	9.5	21%		
18	30	19		Java, Indonesia	Severe rules hedge the sexual life of women prior to marriage in Java, where this female virtuoso produces a tone like a silver wire to a gamelan accompaniment. R (Kunst and Lomax, B 20)
18_1	2 – 6	12	63%		
18_2	5.5 – 9.5	5	26%		
24	38	28		Northeast Thailand	A country girl from this highly stratified, irrigation culture sings in a tense voice indicative of the sanctions and responsibilities that weigh upon S.E. Asian women. Her song thanks Buddha for the beauties of his creations—especially women. Mouth organ accompaniment. (Kaufman #1, A 4)
24_1	3 – 8	21	75%		
24_2	23.5 – 27.5	7	25%		
30	53		100%	Campania, S. Italy	A muleteer, engaged in transport **in** the rugged hills of Campania, sings a traveling song as he rides along, in the high, tense voice so common to the region. (Lomax F.R.)

Figure 4.3.2: The Cantometrics examples for vocal width - summary of the physiological stability analysis. Each coloured block represents a Cantometrics track. The blocks coloured indigo represent tracks which were rated as wide, open and relaxed in terms of Cantometrics vocal width; the blocks coloured yellow represent the Cantometrics narrow and tense ratings; the orange blocks stand for mid/speech tone. We removed the colouring from track 28 while after having performed the physiological analysis we believe that this track was mislabeled in the Cantometrics experiment.

of 17 descriptors / 29 dimensions.

At the planning and design stage of our study a trade-off between the time participants were asked to offer on the one hand and the amount of rating they were supposed to do on the other hand had to be considered. In section 4.2 a trade-off between the number of participants and the number of tracks to be rated was explored. In this section we turn to the ontology descriptors to be included into the rating procedure.

Our aim was on the one hand to chose the subset of the ontology that would cover the vocal production space spanned by our musical examples. On the other hand, we wanted to include as few descriptors into the analysis as would be reasonable in order to allow for more tracks to be analysed within the limited interview time.

Two main reasons were considered to not include descriptors from our ontology into the analysis: the absence of visual information about our musical examples and the time scale (granularity) of analysis. Also, vocal fry and flute register dimensions were excluded because they were not represented in our musical examples discussed in Section 4.3. The resulting subset of the ontology that was used for analysis in our study is presented in Table 4.3.

4.4.1 Visual factor

The main part of our vocal apparatus is concealed within our bodies and cannot be assessed visually without clinical instruments. Yet visual analysis plays a part in evaluating vocal physiology. Singing teachers and medical professionals would always assess their student/patient visually to diagnose either vocal health issues or inefficiency/technical difficulties in vocal production. They would look for signs of strain/effort in the body or face, any involuntary or unconscious movements/tensions; they would examine the posture alignment; also, the visible parts of the vocal tract – lips, jaw, tongue – would be inspected. Possibly also information on larynx height can in some cases be deduced visually. We are not aware of any systematic investigation into the role of the visual factor in assessing vocal physiology.

Our experiment lacked the visual factor: participants were presented with audio recordings of singing with no visual information available.

4.4.2 Granularity/time scale of analysis

There are various considerations that play into the decision on which time scale an excerpt of singing should be analysed. It can range from as little as several milliseconds of sound (like in spectral applications in voice science) to a whole song (e.g. auditions) or even repertoire (musicological studies of a composer or an epoch, comparative studies). Table 4.2 summarises various analytical approaches and the respective time scales. Often the analysis is conducted on a smaller scale and then generalised for larger entities. There are methods to achieve this in MIR such as averaging over sliding window frames to compute MIR features. The advantage of computational, automated solutions is that the number of analysed entities can be large, as long as the process of generalising the results is well defined and justified. In case of manual analysis, like in this study, the number of entities to be analysed has to be small.

There are two contrasting approaches to reduce the number of entities for analysis: either to focus on those that are typical and representative or to bring out outliers, those that are specific for the chosen subset of data. For example Western musicology more often employs the latter approach, studying the most distinguished composers and performers, while ethnomusicology rather stresses the typical, everyday aspect of music making in a culture (though both disciplines nowadays include studies of both types). Cantometrics studies examples of singing which are typical and representative for a tradition.

In Cantometrics a single musical profile is created for a whole culture, based on an analysis of several samples of singing, varying in length from 20 to 90 seconds. This was the scale of analysis Alan Lomax deemed to be suitable and sufficient to rate perceptual descriptors used in Cantometrics. Our aim is to produce ratings on a similar time scale to allow comparison to Cantometrics. Yet we want to refine our analysis and base in on more objective principles, therefore a shorter time frame could be considered. On the other hand, a very short time frame for analysis would result in a large amount of analysis frames. This is not a problem for an automatic approach but where human listeners are involved there is an obvious limit on the number of analysed entities. Also, a very short time frame of several milliseconds would make it impossible for human listeners to analyse auditorily-perceptually complex phenomena like vocal physiology.

For this study we chose several seconds as a time scale for analysis. This time scale allowed us to extract vocal snippets which we considered physiologically stable

Table 4.2: Analytical entities and their temporal scale.

Analytical approach	Time scale
Spectral analysis	20-50 msec
Vowel change	250 msec
Physiologically stable snippets	10 sec
Cantometrics samples	50 sec
Cantometrics dataset – one culture representation	500 sec
All recordings of singing in a given tradition	Hours to millions of hours

according to a definition of a vocal tract setting in phonetics (e.g. Laver 1980, p. 10): a long-term configuration underlying momentary articulations (see Section 4.3). Our snippets are mostly between 4 and 13 seconds long. They allow a more detailed analysis than in Cantometrics, and at the same time the results can be generalised to the Cantometrics tracks.

4.4.3 Choosing the ontology subset

The posture descriptors *head/neck anchoring* and *torso anchoring* from the original ontology (Table 2.2) are mainly identified visually. Though they are very important for sound production and vocal technique, the acoustic outcomes of changing these settings are very subtle and can usually only be heard in a direct comparison. They were therefore excluded from the study.

Articulation descriptors *tongue*, *jaw* and *lips* required special attention. They all change along phonetic segmentation, with every phonetic sound we make, therefore, on a smaller time scale than the length of our snippets. Also, they are assessed visually because they constitute the visible part of the vocal apparatus. These were all good reasons to discard the named descriptors for the study. Yet there can be long-term effects beyond the phonetic changes related to these organs. The three descriptors differ in the significance of the acoustic outcomes associated with their states, in particular regarding longer sung vowels: position of the tongue is defining for the first two formants, while jaw and lips are secondary (see Sundberg 2009); the tongue is also directly connected with the laryngeal area. Also, in terms of visual assessment, the tongue can only partly be seen, thus its assessment always includes an auditory element. We therefore decided to keep the tongue and to discard jaw and lips.

In contrast to all other descriptors which can be assigned to any fragment of singing, onsets are rare events and not every singing fragment would necessarily

contain onsets. The decision was taken to omit onsets from physiological evaluation avoiding the problem of insufficient data at the quantitative analysis stage.

Vocal fry and flute registers were discarded because they were not represented in our musical examples. In accordance with our aim to map the tension component of the Cantometrics *vocal width* parameter (see Section 2.3) a new dimension was introduced to the register categories. We hypothesised that in an analogy to speed gears, the higher the pitch of vocalising in a given register, the more significant would the perception of tension be. So instead of nominal descriptors identifying presence or absence of each of the remaining registers we introduced a 5-point scale for each of them to indicate the part of range for the given register. We expected that this change would have little implications on the rating time.

We had concerns about a possible conflict between the classical registration terminology and the Estill vocal folds vibration mode language. Our worry was that for the participants who do not work with chest/head dichotomy the register descriptor might not be acceptable; yet if they decline to rate it, they would miss the M2 laryngeal mechanism (falsetto) when they talk about vocal folds vibration modes. It was decided to add a nominal dimension to that descriptor indicating presence or absence of falsetto.

Table 4.3 shows the resulting subset of our vocal production ontology that was systematically rated by our participants. The subset is comprised of 11 descriptors and 18 dimensions.

It has to be mentioned that excluding ontological descriptors from the interview protocol did not mean that participants were discouraged from analysing these descriptors: they were prompted to give their opinions on the choice and significance of descriptors. If a participant considered anything not contained in the interview protocol important it was of course recorded and analysed. Excluding descriptors meant that no systematic ratings were collected for them and no quantitative analysis was performed.

4.5 Interview design

To design the interview procedure decisions have to be made what information is presented to the participants in which order and what questions they should be expected to answer in the given interview time. The goal is to elicit as much of their implicit and explicit knowledge and relate it to our ontological space and to the Cantometrics ratings.

Table 4.3: The subset of our ontology of vocal production chosen for analysis in the study.

descriptors	**physiological dimensions**	**range**	**scale**	**metric**
subglottal pressure	subglottal pressure	low to high	5-point	interval
transglottal airflow	transglottal airflow	low to high	5-point	interval
phonation	phonation breathy	present/absent	2-point	nominal
	phonation pressed	present/absent	2-point	nominal
	phonation neutral	present/absent	2-point	nominal
	phonation flow	present/absent	2-point	nominal
register/ position within register range	position within chest register	low to high	5-point	interval
	position within head register	low to high	5-point	interval
	position within falsetto register	low to high	5-point	interval
vocal folds vibration mode	vocal folds modal vs. falsetto	modal/falsetto	2-point	nominal
	vocal folds vibration mode thick to thin	thick/ mixed thicker/ mixed/ mixed thinner/ thin	9-point	interval
thyroid cartilage tilt	thyroid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
cricoid cartilage tilt	cricoid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
aryepiglottic sphincter	aryepiglottic sphincter	wide to narrow	5-point	interval
larynx height	larynx height	low to high	9-point	interval
velum	velum	low to high	5-point	interval
t**ongue**	tongue height	low to high	5-point	interval
	tongue compression	present/absent	2-point	nominal

Vocal space When human listeners judge vocal production they rely on their mental representations of singing: what kinds of vocal behaviour are acceptable as singing, how varied it can be, etc. Their mental representation of singing serves as an internal golden standard (Kreiman et al. 1993) against which new examples of vocalisation are compared. This mental representation defines for each physiological or other parameter how it usually sounds in respect to it, what deviations are possible and how extreme they can be. Given a set of parameters with which the rater is familiar we can talk about an inner vocal space spanned between the rater's inner representation of the parameters' most extreme values.

This inner vocal space is conditioned in three ways: a) by the rater's experience as a listener (perceiving, analysing, judging singing), b) by the rater's experience of vocalising (which sounds are easy/possible, how are they produced, see Section 8.2.2.7) and c) by the cultural context, in particular, cultural preferences about singing. The cultural context determines to a large extent what kinds of singing a person listens to and what kinds of sounds they make; therefore its influence on the person's vocal space is crucial.

No anchoring Because everyone's vocal space is different, it is sometimes useful to anchor or equalise the raters' spaces giving them examples of scales extremes and mid (zero) positions. Alan Lomax provided this kind of training for Cantometrics raters in his book accompanied by audio cassettes specifically released for this purpose (Lomax 1977). While such a training helps to bring the raters to a better consensus and make their ratings comparable, it applies a filter on their knowledge (a transformation of their vocal space). The use of such a filter and in particular the choice of examples on which the raters are trained should be well justified.

We decided not to include this pre-training stage in our experiment. In contrast to Cantometrics, we were planning to interview experts with decades of experience in vocal production. We were interested in eliciting their knowledge and experience, without any kind of filters. That could possibly lead to a better understanding of the interpersonal or even intercultural vocal space. Introducing pre-training would confine our results to the vocal space spanned by the training examples. Who would be the authority to decide about these examples? Our participants would be more experienced in vocal production then us or the Cantometrics team.

Order of presentation Given the expertise of our participants we assumed that their understanding of the descriptors would generally be very good, therefore no

randomisation of the descriptors order was necessary.

Yet we expected that they will be unfamiliar with the majority of the musical cultures represented by our musical examples. Therefore we could not exclude that listening and analysing singing in an example that might seem exotic to them would not influence their vocal space and thus the further analysis. For that reason tracks should be presented to them in a randomised order. In Section 4.2 we demonstrated that with high probability each track would be analysed sufficiently frequently.

We wanted to collect perceptual ratings alongside physiological in order to relate the results to the Cantometrics ratings. Yet we assumed that performing detailed physiological analysis of a singing sample might affect the perception of it. We therefore decided to ask the participants to rate perceptual descriptors before and after the physiological analysis.

No restrictions were placed on participants hearing the tracks in advance as well as on giving them information on the origin of the tracks.

Defining the terms Since no training phase was planned in the interviews the definition and explanation of terms would have to be integrated with the analysis of the first track.

We expected that defining the terms exactly would be difficult, in particular because our participants will have their own understanding of them. We could force them to use our definitions presented in Chapter 2, yet that would have a number of negative consequences. Some of them might be unfamiliar with the language used in these definitions. Others might object to that language or to the definitions and their cooperation would be diminished. Insisting on our definitions would prevent us from eliciting their own use and understanding of the terms. We therefore decided to loosen the approach and rely on our participants' previous knowledge. In case the terms and the language were unfamiliar, we would give explanations using synonyms or vocal function.

Further data Because physiological analysis is a complex process, and results can be ambiguous we introduced the confidence ratings. Alongside each physiological rating our experts would provide their confidence in that value.

Another aspect we wanted to document was the salience of physiological descriptors for the perceptual questions we asked. Apart from it being an important research question in itself, we envisioned using this information for triangulation. The question on salience would be asked at the very end of analysis of each track.

Opportunities for discussion Our mixed-methods study was planned to contain both quantitative and qualitative analysis. To facilitate the quantitative analysis, the interviews has to be structured and participants' answers comparable. At the same time, because the outcome was completely open, to make the qualitative analysis worthwhile it was important to leave enough space and opportunities for discussion during the interview. Therefore we intended to prompt participants to comment on each new term they encountered, on the process of analysis, on any difficulties or ambiguities they were aware of, to make suggestions for terms or for other aspects of the experiment design. To ensure there is always enough time for discussion we decided no to limit the time spent on each track. That made the actual coverage of the tracks in terms of analysis less predictable, but we decided that analysing what experts had to say was more important.

Interviewer It is important here to mention the interviewer and their role during the interviews. Interviewer's biases, judgements and behaviour have a significant impact on the interviews and should be taken into account.

In our case the interviewer, the author of this thesis, was well-versed in the terminology proposed in the ontology. She was trained in the Estill model (did the Estill level 1 and level 2 course), and took part in Sundberg's voice science Summer School. She was an experienced singer and musical director, having been on the receiving and on the giving end of singing tuition. She therefore could speak the language of the models suggested for analysis, but could, if necessary, adjust her language to other backgrounds.

The interviewer also had an ethnomusicological background. She conducted field research in various regions of rural Russia, provided extensive annotations, conducted a study on the freedom of musical expression for UNESCO, worked extensively with ethnomusicological archives. As a singer she had been involved in a large number of vocal traditions. Through these activities she acquired a varied experience in cross-cultural vocal production, which presumably reduced the effect of her cultural bias.

Interview protocol The resulting interview protocol consisted of three phases: 1. Background, 2. Physiological and perceptual analysis and 3. Wrap-up. At the beginning of the interview information about the participant's background, biography including in which cultures they lived, their education, musical and singing experience, profession, how often and in which circumstances they deal with vocal

physiology, etc.

At the next stage the participant was first presented with a track and asked about their perception of it in the Cantometrics related oppositions of wide/narrow and tense/relaxed. Then a snippet of the track was played back, after which physiological analysis began. Each descriptor would be rated, rating discussed and a confidence score provided by the participant. The researcher would prompt the participant to express any doubts, concerns about the descriptor or the rating, suggest alternatives. After all descriptors were rated the next snippet (if available) was handled in the same way. When all the snippets were analysed physiologically the participant was pointed back to the perceptual characteristics of the track and asked to make changes to their original judgements if necessary. This information provides us with unique data on how detailed physiological analysis influences our perception of singing. The participant was then asked which physiological descriptors were salient for their perception of the track and which were unimportant. This generated another unique set of data with judgements about salience of physiological aspects; it could provide triangulation at the dimensionality reduction stage of the analysis. To conclude the analysis of a track the participant gave their opinion on the representativeness of the snippets.

Then the next track was presented and its analysis followed the same steps, and so on. In the first round the researcher made sure that the participant understood the ontological terms and gave explanations where necessary. In the last wrap-up phase the participant was once more asked about feedback on physiological descriptors and further suggestions. Then the question of descriptor salience was generalised for all tracks. The last minutes were used to discuss anything that the participant felt was left out.

4.6 Data collection

While compiling a pool of participants for our study, we had to take following into consideration:

- all participants should be experts in the field of vocal physiology
- they should be diversified in terms of their profession / field of study
- ideally, some diversification of cultural background and experience should also take place

Table 4.4: Diversification of participants

Total interviews	13
Medical professionals	6
Singing teachers	8
Scientists	6
Influence: Jo Estill approach	5
Influence: Complete Vocal Technique	2
Influence: Johan Sundberg	4
Non-Western musical background	3

- also, known methodological influences (such as the Estill system) should be taken into account

We managed to secure support of 15 professionals who agreed to take part. All of them have been involved with vocal physiology for a long time (10-45 years) and represented a variety of fields and occupations (Table 4.4). Of 13 experts who were interviewed during our study, six were medical professionals (larynx surgeons, phoniatricians, speech and language pathologists), eight were singing teachers, covering a large variety of contemporary Western singing styles (classical, pop, rock, gospel, jazz and more) including one teaching a non-Western tradition. All participants belonged to one of these two categories, with one belonging to both. Six participants were actively involved in scientific research about singing voice (three of them medical professionals and four singing teachers).

Five participants admitted to having been familiar and influenced in their work by Jo Estill's approach, though all of them mentioned that they do not use her system in its entirety or have moved away from it at some point in their career. Two participants belonged to the Complete Vocal Technique school founded by Catherine Sadolin (Sadolin 2000), which is remarkable for its approach to teaching vocal effects and non-conventional vocal production techniques. Four experts were closely linked to Johan Sundberg and his research.

Unfortunately, it was very difficult to find vocal physiology experts from cultures other than Western, therefore the diversification of cultural background was only realised partially. We managed to recruit three singing teachers with strong links to non-Western traditions (to South Africa, Russia and Indonesia), though only one of them currently lives outside the Western world. We were not able to locate any vocal physiology experts among ethnomusicologists. Interviews took place September through November 2013. Of 15 planned interviews only 13 took place: one

expert had an unexpectedly high workload and couldn't spare 90 minutes; another interview failed due to technical difficulties. Most interviews were conducted via Skype/internet calls, only in two cases were the researcher and the expert present in the same room. Prior to the interview participants were sent a consent form, a link to download their personal (compiled in a random order) playlist of musical examples and a physio analysis form. They were asked to read carefully and to fill out and sign the consent form and send it back to the researcher. They were also asked to download the playlist and make sure that they had a technical setup to comfortably listen to and analyse the music during the interview. Participants were not required to listen to the examples prior to the interview, but there were no restrictions on doing so either. The physio analysis form had to be printed out to be used (discussed and filled out) during the interview.

There were no risks to the participants due to the study and an ethical approval was received from the authors' institution. The main concern of the consent form was the storage and the preservation duration of the interviews' recordings. The interviews were recorded as audio for qualitative analysis purposes. Since it became obvious that the authors won't have the resources to transcribe all the interviews, the question of retaining the recordings to enable experiment replication at a later date arose. The consent form made participants aware of the fact that the interviews were recoded and explained the details of recordings' preservation and access. There was also an option to change the sound of the voice in the recording to avoid participants being recognised – though none of the participants opted for it. We received the signed consent form from 12 out of 13 interviewed participants. The recordings of the interviews with those 12 participants are stored in an electronic archive[2] available to future researchers for any research purposes. Anonymised quantitative data is publicly available; for accessing interviews and consent forms permission has to be sought.

For the researcher conducting the interview the preparation included creating a personalised playlist for the participant; sending the above mentioned documents to the participant with an email explanation; answering any questions participants had about the consent form or the interview; making sure that the consent form was filled out and sent back; printing out an interview protocol template and a physio analysis form to be used during the interview. During the interview the technical

[2]Open Science Framework is a free and open source research project management system which guarantees preservation and controlled access. The current project can be found at `https://osf.io/pff8m/`

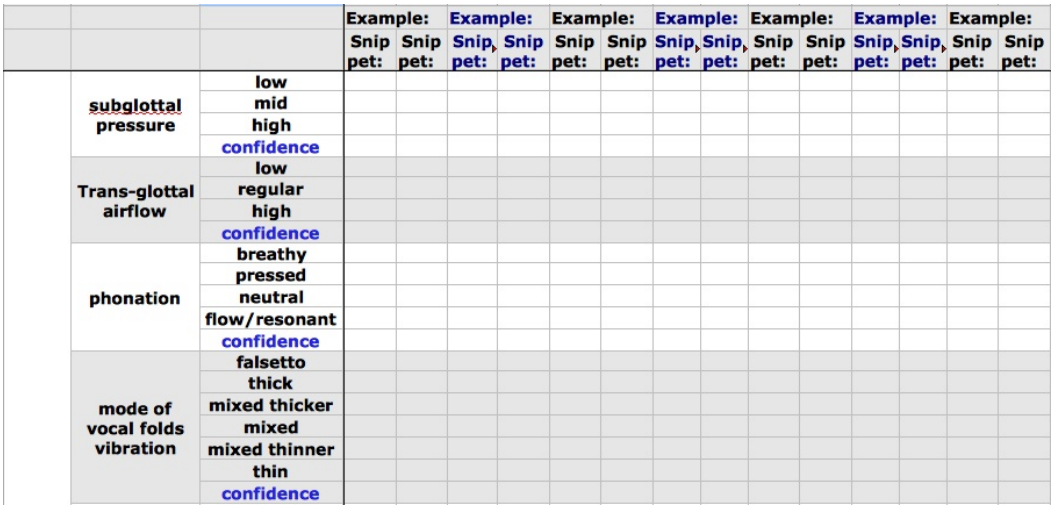


Figure 4.6.1: A fragment of the template for physiological analysis

setup included a voice-over-IP/Skype connection with the interviewee, an audio recording software capturing the interview, an external audio recording device to ensure a backup recording in case of a computer failure, an audio playback software (e.g. iTunes) to sample the musical examples if necessary. The researcher used the interview protocol template to keep track of the interview structure and to make notes about the participant's answers to open questions, remarks or criticism. The physio analysis form was used to mark the participant's physiological ratings: each rating was discussed (sometimes in a great detail), and both the researcher and the participant filled out their exemplar of the physio form. This simultaneous use of two exemplars of the form was a practical solution that allowed the participant to review and revisit the results of their analysis as well as the researcher to have the data right there without a need to request the data over the internet.

Quantitative data was collected in the form of participants' ratings of physiological descriptors from the chosen subset of our ontology (Table 4.3). Each participant rated only those descriptors, which they were familiar with or could make sense of after an explanation given by the researcher. The participant rated as many tracks as the time of the interview allowed. While 90 minutes were usually planned, some participants were happy to continue beyond that limit or to schedule an additional interview.

Together with the ratings of physiological descriptors participant's confidence in these ratings was recorded. This was desirable, since in many cases interviewees could not provide a completely unambiguous and reliable rating, due to not being

Table 4.5: Example of quantitative data collected during an interview

physio	type	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
subglottal pressure	R	3	1	3	3	1	1	1	2	-1	3	5	-1	-1	-1	1	1	-1	-1	-1
subglottal pressure	C	3	4	3	3	5	4	4	4	-1	3	3	-1	-1	-1	3	3	-1	-1	-1
transglottal airflow	R	3	3	3	1	1	1	3	1	-1	1	3	-1	-1	-1	1	1	-1	-1	-1
transglottal airflow	C	3	4	3	3	5	4	4	4	-1	2	3	-1	-1	-1	3	3	-1	-1	-1
phonation breathy	R	0	1	-1	-1	0	0	0	0	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1
phonation pressed	R	0	0	-1	-1	0	0	1	0	-1	0	1	-1	-1	-1	-1	-1	-1	-1	-1
phonation neutral	R	1	0	-1	-1	1	1	0	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
phonation flow	R	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
phonation	C	4	4	-1	-1	5	5	4	4	-1	5	4	-1	-1	-1	-1	-1	-1	-1	-1
vocal folds falsetto	R	0	1	0	0	0	0	0	0	-1	0	0	-1	-1	-1	0	0	-1	-1	-1
vocal folds thick to thin	R	2	0	3	3	5	4	1.5	4	-1	5	5	-1	-1	-1	3	3	-1	-1	-1
vocal folds	C	3	4	3	3	5	5	5	4	-1	5	5	-1	-1	-1	3	3	-1	-1	-1
larynx height	R	4	2	3	3	5	5	5	3	-1	5	5	-1	-1	-1	2	2	-1	-1	-1
larynx height	C	4	4	3	3	5	5	3	5	-1	5	5	-1	-1	-1	5	5	-1	-1	-1
thyroid cartilage tilt	R	3	1	5	5	5	3	1	3	-1	5	3	-1	-1	-1	5	5	-1	-1	-1
thyroid cartilage tilt	C	4	4	5	5	5	5	4	5	-1	5	5	-1	-1	-1	5	5	-1	-1	-1
cricoid cartilage tilt	R	1	1	1	1	1	1	1	1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1
cricoid cartilage tilt	C	5	5	5	5	5	5	5	5	-1	5	5	-1	-1	-1	5	5	-1	-1	-1
velum	R	1	1	1	1	5	5	1	1	-1	3	3	-1	-1	-1	1	1	-1	-1	-1
velum	C	5	5	5	5	5	4	5	5	-1	3	3	-1	-1	-1	4	4	-1	-1	-1
AES	R	3	1	3	3	5	5	3	3	-1	5	5	-1	-1	-1	3	3	-1	-1	-1
AES	C	4	3	4	4	5	5	5	4	-1	5	5	-1	-1	-1	4	4	-1	-1	-1
tongue height	R	0	3	3	3	2	3	3	1	-1	2	2	-1	-1	-1	0	0	-1	-1	-1
tongue compression	R	1	0	-1	-1	0	0	0	0	-1	0	0	-1	-1	-1	1	1	-1	-1	-1
tongue	C	5	4	5	5	5	5	5	4	-1	5	5	-1	-1	-1	5	5	-1	-1	-1
position within chest register	R	3	0	3	1	0	4	1	3	-1	0	0	-1	-1	-1	1	3	-1	-1	-1
position within head register	R	0	0	0	0	1	0	0	1	-1	2	5	-1	-1	-1	0	0	-1	-1	-1
position within falsetto register	R	0	1	0	0	0	0	0	0	-1	0	0	-1	-1	-1	0	0	-1	-1	-1

able to see the singer, a non-optimal recording and playback quality and generally because physiological mechanisms are not always well understood and do not always lend themselves well to this kind of analysis. Confidence was rated on a 5-point Likert scale with 5 being absolutely confident and 1 meaning a random value.

All the data collected for the study can be found in the Open Science Framework repository: `https://osf.io/pff8m/`. The project is called "Vocal production ontology" and contains a number of components with different access permissions. We chose Open Science Framework because it is widely recognised in the scientific community and data repositories on OSF are accepted by many scientific journals; it also offered a clear and easy-to-use structure that allows for long-term data preservation and flexible sharing with various levels of access. The Interviews component of the project contains audio recordings of the interviews, interview transcripts where available and consent forms. These are the documents which allow direct or indirect identification of the participants. These documents will remain "private" and an unlimited access to them will be granted to the participants only. Researchers wishing to access these documents and media will have to obtain permissions from the participants. Musical examples, methodological documentation as well as quantitative data that is not linked directly to participants' names and personal data will be openly available to facilitate experiment and analysis replication.

Table 4.6: Which tracks were rated by which participants. Each snippet was rated by at least five participants. Exceptions are snippets 22_2 and 32_3. The former refers to the lower part in snippet 22_1, and was rated only by those participants, who, after rating 22_1 explicitly wished to provide an alternative rating for the lower part. The latter was not originally chosen by us for analysis, but was suggested by a participant; it was only presented to those participants who expressed doubts about our choice of snippets in this example.

Tracks		16	18	20	22	24	26	28	30	32	34
P1		1								1	
P1_1	1		1	1	1	1					
P2		1	1	1	1	1			1		1
P3											
P4			1	1			1	1		1	1
P5			1		1	1		1			1
P6											
P7	1		1	1		1				1	1
P8			1					1			1
P9					1		1	1		1	
P10	1	1	1	1	1	1	1	1	1	1	1
P11	1				1	1			1		1
P12						1			1		
P13	1	1		1	1			1		1	1
P14	1	1	1	1	1	1	1	1	1	1	1
P15	1	1			1	1	1		1		
P16											
Total											80

4.7 Analysis

The analysis should follow the steps listed at the beginning of this Chapter. Each next step depends on the success or otherwise of the previous one. If ontology terms are well accepted by the experts then inter-rater agreement can be investigated. Only if consensus is demonstrated can annotations be collected. If consensus cannot be confirmed the main task becomes to explore common themes and confounding issues to gain insights about reasons for disagreement.

In this section we discuss how ontology acceptance can be measured quantitatively or inspected qualitatively. We then discuss inter-rater agreement and various considerations which need to be taken in account. Depending on the outcome of that step we suggest two ways to continue the study. We round up with an account of the actual iterative development of the study, decisions which had to be taken and their consequences.

4.7.1 Ontology acceptance

The first step in data analysis is to estimate the acceptance by the participants of the ontology terms and of the overall study design. Because participants were free to leave out the analysis of the terms they were not familiar with or did not agree with, a quantitative measure of acceptance for each descriptor would be the proportion of the participants who were happy to rate it. Qualitative analysis would reveal participants' views on the appropriateness of the terms: whether they are relevant to the analysis; possible pitfalls and ambiguities such as polysemy or biases related to particular genres/communities; suggestions of better alternatives. The confidence and salience questions could provide further conformation. All the above concerns not only the terms but also the scales they are rated on. Additionally, participants may propose new terms not included into our ontology.

If any serious failings regarding the ontology are picked up along the way during interviews, the qualitative component of the study could be modified and continued in an iterative way. The quantitative component, if affected, might be compromised. Inappropriate terms can be dropped without damage, yet adding new ones after some interviews have already been conducted would make the data unbalanced and the judgement on the acceptance of the new terms unreliable. Changing terms/scales in the middle of the experiment would render quantitative data on them meaningless. Therefore the quality of the ontology terms is decisive for the success of the experiment.

4.7.2 Inter-participant agreement

The next step to take is to establish the presence of an inter-subjective consensus on the meanings of the ontological terms. This can be achieved through statistical analysis of the participants' ratings. Various measures of agreement and/or reliability can be applied (see Kreiman et al. 1993 for more details). Agreement refers to direct equivalence of the ratings: the less the distance between the two ratings the higher the agreement. Reliability (also called inter-rater consistency) takes in account a possible constant bias: two ratings are reliable/consistent if the distance between them remains constant for different observations or if their functions are correlated. The two approaches measure different qualities: perfectly consistent ratings can display poor agreement and good agreement does not guarantee consistency. In our case, because the scales we used only had a small number of states, we decided that agreement was a more appropriate concept.

There are several widely used measures of inter-rater agreement including Fleiss' κ for nominal dimensions, Pearson's r or Spearman's ρ correlation coefficients for interval dimensions or intra-class correlation coefficient ICC. An important consideration here is that there will be missing data – not every snippet was rated by each participant. We chose Krippendorff's α as a modern statistic that generalises all the above measures, is applicable to both nominal and interval ratings and can handle missing data (Krippendorff 2012).

Another factor that needs to be scrutinised is the intra-rater consistency – a measure of how similar a rater would judge the same stimulus (analyse the same snippet) on e.g. different days. It is common in studies like ours to present the raters with the same stimuli to measure their consistency over time. In our case the participants' availability and time were very limited. We decided against presenting them with the same tracks, thus saving more time for new ratings and discussion. We argued that our participants were experts who had been dealing with various kinds of vocal production routinely for at least a decade; therefore their mental representations of the vocal space (possible instances of singing) were saturated and would not be affected by our examples. A common technique to increase intra-rater consistency is to train them in the rating procedure with stimuli examples; our experts were trained for objectivity in decoding physiology through their daily work and thus the risk of drift in their ratings would already be reduced. We deliberately refused to subject our participants to any rating training procedure apart from an open discussion of the ontology terms in the first round of analysis. The reason for

this decision was the fact that any training we could devise would have been biased by our knowledge and experience of vocal production which was inferior to that of our experts. We were interested in our participants' views unaffected by our ideas.

It is important to remember that a good agreement statistic can be a result of a coincidence, not consensus. It is therefore essential to check for statistical significance by calculating the confidence interval for the statistic. If zero falls within the confidence interval the chance of agreement being the result of coincidence is too high. If agreement is high and the confidence interval lies way above zero a consensus between raters on the descriptors' values has been demonstrated.

In this case the (averaged) ratings we collected could be attached as reliable annotations of vocal production to the 19 musical snippets that were used for analysis thus comprising the first cross-cultural dataset of singing with vocal physiology/vocal production annotations. By way of generalisation, our method – eliciting tacit knowledge of experts about vocal physiology – could be applied to annotate further datasets of singing recordings and to construct the so-called ground truth to be used for training of computational models in machine learning, which, if successful, would allow to increase the number of annotated tracks significantly through automated procedures.

If participants do not agree about the ratings, no intra-subjective consensus can be claimed. In this case the main interest of the study becomes to analyse the reasons for disagreement and the qualitative component becomes crucial: a detailed analysis of the answers to open questions, of participants' concerns and misunderstandings, of problem cases might provide us with insights into the underlying reasons for differences in ratings. If these reasons appear to be structural and pervasive and not due to special cases or technical failures it could be argued that given the current state of knowledge vocal physiology is not suitable for objective annotation; that under conditions similar to those of our study even highly experienced experts cannot reliably decipher vocal physiology from audio recordings of singing. This claim would have wide ranging implications for various voice-related disciplines, in particular for singing education and for MIR.

4.8 Study progression

The acceptance of the ontology terms was good (Section 5.1), the only descriptor that participants were often unfamiliar with was *flow* phonation (Section 5.2). We therefore could proceed to the inter-participant analysis.

Given the complex and fuzzy nature of our subject we expected fuzzy inter-participant agreement results, with some descriptors showing somewhat better agreement and others rather less. Yet our results were more decisive, with good agreement on two descriptors (*AES*, *larynx height*), a tendency to agreement on the third (*subglottal pressure*) and no agreement on all other descriptors (Chapter 5). This result clearly signalled the preference for the second option to progress with the study – qualitative analysis, searching for underlying reasons for disagreement.

The workflow for the qualitative analysis of the interviews was planned as follows:

1. create structured codes for known categories beforehand
2. listen to everything twice. At the first listening:
3. write a précis of the interviewee's background
4. segment the audio
5. code all relevant concepts and vocabulary by means of open coding along the way
6. write memos
7. code good citations
8. code interesting concepts that don't fit into our ontology
9. code validation of our experiment design
10. create a raw code structure
11. Listen for the second time, check the codes and the segmentation, make sure that nothing of importance has been missed, add or restructure as necessary.
12. Assess whether the main goal – extracting common themes and confounding issues – has been reached and what insights have been won.
13. If appropriate, conduct ontological analysis and structuring of the codes, possibly in a different software.

These procedures defined the constraints for the analysis framework. The first challenge we faced was to find a suitable software package that would adhere to these constraints. The requirements were set as follows:

1. can work well with audio sources
2. allows transcription in place
3. allows audio segmentation in place, convenient
4. optional: import of external segmentations and transcriptions
5. is convenient for open coding
6. provides sensible visualisation for the codes, their structure and their occurrence in the sources, filtering
7. memos
8. optional: export of transcription with time stamps
9. tools for structuring codes
10. exporting codes
11. optional: ontological analysis of the codes, mapping relationships
12. optional: mixed methods analysis with relation to the ratings
13. good support
14. cheap or licence available

The choice of the software was crucial because it would largely define the analysis procedure and determine the day-to-day work for the researcher. The analysis software is the research medium that shapes the view of the data and the ways to filter and present it.

We tested several software frameworks for qualitative analysis according to the constraints listed above. We chose a widely used qualitative analysis software NVivo: it met most of the requirements on our list. In particular, it was more stable than most other products, offered good segmenting facility, comfortable coding, great visualisation and reliable support. The main disadvantage of NVivo is that it is a proprietary, very expensive product. While the university provided a license for the time of the PhD, it is to be assumed that the researcher will not be able to revisit the codes in future without access to NVivo. Neither would other researchers be in a position to replicate the results without an NVivo license.

Also the question of transcribing the interviews had to be dealt with. It is common to transcribe interviews for analysis; transcriptions offer many advantages since text is much easier to handle than audio: direct search, occurrence statistics and more. Yet transcribing is very time-consuming and outsourcing it would have a cost. There was no budget for transcriptions and transcribing everything was not an option due to time constraints. We chose a balanced solution: some participants expressed themselves very clearly, using well-structured sentences, conveying their thoughts in logical, didactic manner; others mumbled, spoke in half sentences, thought aloud, etc. It was much harder to extract the essential information from the latter interviews based only on the audio. We therefore commissioned the latter interviews for transcription (3 interviews out of 13). For all the other interviews we only transcribed the bits that were important for further analysis, such as discussions of study design, physiological parameters, salience, etc. This transcription was done on the fly during the segmentation and coding process.

The main issue in the process of analysing the 33 hours of interview recordings was the time. Qualitative analysis took much longer than originally planned, we hugely underestimated the time needed for it. While progressing slowly with the analysis we searched for ways to narrow the scope of the analysis in the way that would still allow to make conclusions about our research question – the underlying reasons for disagreement between experts. We suggested to concentrate our analysis mainly on one track: Track 24 was rated by many participants and offered a wide range of disagreements (Section 6.1). Yet further into analysis it became clear that this choice would be too limiting for further conclusions and we widened our scope. Track 24 provides a good focus to begin investigating the qualitative data, and we kept it as a starting point for our presentation of the data, which begins with a strong focus on it and then gradually widens the description.

Another attempt was to choose two descriptors that are very different and represent a variety of views and attitudes. We chose *larynx height* – one of the two descriptors which showed good agreement, and *velum position* – a descriptor we expected experts to agree about, but no agreement was found in the quantitative analysis. Our hope was that looking at these descriptors and contrasting participants' views and behaviours about them we would see patterns of agreement or disagreement. Soon it became clear that there were too many specificities regarding each of the descriptors that would make generalisations based on observing just these two unjustified. For example, there were difficulties with defining both of the agreed descriptors – *larynx height* and *AES* – but these difficulties were of different

nature for each of them, therefore concluding the findings about one of them for the other would not be appropriate. We also found that velum position ratings for one particular participant were contrary to the most other ratings; we assumed that the participant rated nasality instead of velum position, therefore inverting the ratings. This was the only such case in our study. We tried inverting his ratings, but still no agreement could be claimed. Eventually we decided against limiting our analysis in terms of descriptors.

In Chapter 6 we present representative excerpts from the interviews and an initial analysis for each descriptor. Some presentations are focussed on Track 24, where we felt that this example gives a good overall understanding of disagreement issues; others include further tracks.

Open coding is very convenient in NVivo and was performed on the go while reading the transcript or listening to the recording. For most interviews it was processed in two rounds: the first round picked up the most obvious candidates during segmentation; in the second round more in-depth analysis of concepts and vocabulary took place for the more significant bits of the interviews. There was an iterative element in this process: some concepts and terms occurred in several interviews, and this in itself was an indication of the importance of the concept. Therefore, if a code was created for one interview, we would scan other interviews for related concepts and terms and code them as well.

Open coding provided us with sufficient input to analyse the acceptance of study design and ontology terms (Section 5.1). It also gave us a list of terms suggested for the rating procedure with all the contextual information, its analysis is presented in Section 6.9. Following data presentation in Chapter 6 we performed a meta-analysis, extracting and consolidating six themes related to disagreement that came up in several interviews and in the context of multiple tracks (Chapter 7). These themes represented the kinds of bias the raters were subject to and the main confounding issues that affected our experiment. In Chapter 8 we discussed the consequences of these findings for MIR and for singing education and devised a new roadmap for revising Cantometrics.

5 Interviews – quantitative analysis

In the previous chapter a methodology for the integrated approach to automatic labelling of our ontological parameters was laid out, in particular, a detailed presentation of a mixed-method study aimed at eliciting expert knowledge on vocal production (Chapter 4). The purpose of the study was given as follows:

a) verify the applicability, relevance and completeness of our ontology and adjust it if necessary,

b) examine the consistency of experts' ratings, their agreement about the meanings/values of the descriptors,

c) collect reliable ratings

This chapter describes the statistical analysis of the quantitative data gathered during interviews with vocal physiology experts for the aforementioned study. The inter-rater agreement is calculated for a chosen subset of our ontology descriptors (see Table 4.3) using *Krippendorff's alpha*. We implemented Prof. Krippendorff's bootstrapping algorithm in a popular statistical framework *R*, with an additional introduction of weighted observations and made it available on GitHub to the open source *R* community[1]. Using this routine we then calculated the alpha values and the confidence intervals for all descriptors. The results displayed some irregularities – only two of the descriptors showed a clear tendency to agreement. We performed various checks to verify the statistics were calculated correctly. While the calculations appeared sound, it became clear that Krippendorff's alpha is biased when dealing with sparse data.

We collected average ratings for the two descriptors for which consensus among experts was demonstrated – *AES* and *larynx height*. These are the first reliable annotations of vocal production for a highly varied, cross-cultural dataset. We also demonstrate that both *AES* and *larynx height* correlate with the Cantometrics *vocal width* ratings. We have therefore found a more objective alternative to the

[1] kripp.boot() package and function, `https://github.com/MikeGruz/kripp.boot`

Cantometrics *vocal width* parameter that will retain the correlations to societal traits discovered in the original Cantometrics experiment.

This chapter wraps up with a discussion and future research suggestions, in particular a proposal of a follow-up experiment on experts' consensus on physiology that excludes cross-cultural variation.

Qualitative analysis and meta-analysis extracting possible reasons for disagreement about all other descriptors are dealt with in the following chapters.

All data generated from interviews, the R code and the calculations can be found in the Open Science Framework repository: `https://osf.io/pff8m/`.

5.1 Acceptance of the study design

During our interviews with experts in vocal physiology, we informed the participants about the details of our study design and asked for their opinions. They had lots of opportunities to suggest changes and offer critique and recommendations, encouraged through direct questions and prompts as well as while answering open questions.

Of 13 participants 11 (over 80%) were happy to rate 80% or more of physiological descriptors we chose for rating from our ontology (see Table 4.3). This indicates that our descriptors were generally understood among experts – either previously known or acquired through explanation during the interview.

Experts were explicitly asked to reflect on the chosen physiological descriptors, assess their appropriateness and relevance, to suggest changes to the ontology. A considerable amount of time during the interviews was dedicated to that and the experts' answers to free-form and open questions are analysed in Chapter 6. There was just one descriptor that experts were rarely acquainted with: the *flow* phonation (see Section 2.1.1.1). Some critique was also expressed of the *thyroid* and *cricoid tilt* descriptors (these were the least rated ones after flow phonation) and some differences in describing the position of the tongue were noted. Suggestions of additional descriptors or changes to the ontology are discussed in Section 6.9.

We explicitly asked the experts whether the snippets we presented them for physiological analysis were representative of the whole track, which they heard and were asked questions about prior to the analysis. In almost all cases our choice was confirmed to be appropriate. Only for track 32 there were two participants who criticised the choice of snippets and chose to rate an additional fragment of the track (see Figure 4.6).

Overall, the experts expressed support for our approach which allowed us to proceed to the analysis of the collected quantitative data.

5.2 Reshaping quantitative data

During the interviews we collected two sets of three-dimensional quantitative data: the ratings and the confidence values. The dimensions are:

- participants (13)
- snippets: representative fragments of dominant physiological states present in the musical examples (19, see Table 4.3.2)
- physiological dimensions from our ontology (originally 18, see Table 4.3)

There is a certain sparseness/asymmetry to the data. While participants rated all the descriptors to any track they listened to, they didn't usually get to rating all the tracks, therefore, for any given descriptor, they didn't provide ratings for every track. It can be conceived visually: in the original tables of physiological interview protocols, some columns are filled out completely and others are completely empty (see Figure 4.5).

As mentioned above, flow phonation was largely unknown to participants and only rarely rated. There was therefore not enough data gathered for this descriptor and it was excluded from analysis. Ratings for phonation modes were reformulated to consist only of three classes: breathy, pressed, neutral, and the original flow phonation mode to fall into the neutral category.

A complex descriptor *vocal folds vibration mode* in its original form presented a problem for statistical analysis: vocal folds either vibrate in falsetto (M2) or in the modal mode (M1), and if they are in the modal mode, their thickness can be characterised linearly (from thick to thin). The original descriptor *vocal folds vibration mode* was represented by two dimensions: *vocal folds: modal vs. falsetto* (nominal) and *vocal folds vibration mode thick to thin* (interval, see Table 4.3). The first of the two dimensions separated out the information about the laryngeal mechanism, and the second dimension originally only contained information about the vocal fold thickness. There was a contradiction though, which affected the value given to the new *vocal folds vibration mode* descriptor in case when the snippet was rated to be sung in falsetto and didn't have a thickness value: *NA* was reserved for situations when participants didn't rate the snippet, and therefore couldn't be

used; and any numeric value became a part of the linear scale. Thus this descriptor could not be made independent from *modal vs. falsetto*. In physiological terms in falsetto only the edges of vocal folds take part in vibration. This kind of vibration often takes place with only a partial closure (while the air stream bypassing the folds can be audible or not). We chose a linear solution that we thought would be a closest approximation, labelling falsetto as "thinner than thin" vocal folds: while the thickness of vibrating vocal folds was rated on a 9-point scale from 1 to 5 in 0.5 steps, we assigned the snippets in falsetto the value 9. This way, in the interval metric, confusing thin fold with falsetto was penalised by the same amount as confusing thick fold with thin fold; confusing falsetto with thick folds was penalised twice as much.

We were then left with 17 physiological dimensions and for one of them (*vocal folds vibration mode*) the scale was changed (Table 5.1).

To address the main goal of our study – assess the viability of physiological approach to modelling vocal production – an analysis of inter-participant agreement was paramount. We needed to investigate whether participants rate physiological descriptors similarly. If they did, it would have been a good indicator that, though physiological processes in singing are not completely understood and modelled, experts' tacit knowledge can be used reliably to describe vocal production. If participants did not agree, it would raise many questions:

- are there differences between professional groups
- can physiological information be extracted entirely through listening, without a visual source
- are there multiple physiological configurations that could lead to similar acoustic results
- does the cultural and musical background of a listener influence their physiological judgements
- where are the limitations of our knowledge of vocal physiology

To analyse inter-participant agreement the data was reshaped to the form, where all results on a particular descriptor were collected in one table (Table 5.2), with dimensions *participants*snippets*.

Also, for a better comparability, results were normalised to the range between 0 and 5, so that descriptors with scale exceeding 5 points would have non-integer

Table 5.1: Ontology adjusted

descriptors	**physiological dimensions**	**range**	**scale**	**metric**
subglottal pressure	subglottal pressure	low to high	5-point	interval
transglottal airflow	transglottal airflow	low to high	5-point	interval
phonation	phonation breathy	present/absent	2-point	nominal
	phonation pressed	present/absent	2-point	nominal
	phonation neutral	present/absent	2-point	nominal
vocal folds vibration mode	vocal folds modal vs. falsetto	modal/falsetto	2-point	nominal
	vocal folds vibration mode thick to thin	thick/ mixed thicker/ mixed/ mixed thinner/ thin	17-point	interval
larynx height	larynx height	low to high	9-point	interval
thyroid cartilage tilt	thyroid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
cricoid cartilage tilt	cricoid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
velum	velum	low to high	5-point	interval
aryepiglottic sphincter	aryepiglottic sphincter	wide to narrow	5-point	interval
tongue	tongue height	low to high	5-point	interval
	tongue compression	present/absent	2-point	nominal
register/ position within register range	position within chest register	low to high	5-point	interval
	position within head register	low to high	5-point	interval
	position within falsetto register	low to high	5-point	interval

Table 5.2: Example of reshaped data for the descriptor *subglottal pressure*. It was rated on a 5-point Likert scale. Since each participant only rated a random subset of the snippets, many cells remain empty (carry an NA value). Ratings are highlighted in blue and via a background shade of the cell, which also indicates the rating value.

	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
P01	3	1	3	3	1	1	1	2	NA	3	5	NA	NA	NA	1	1	NA	NA	NA
P02	NA	NA	3	5	3	1	1	3	NA	5	5	5	NA	5	NA	NA	NA	3	5
P04	NA	NA	NA	NA	5	3	1	NA	NA	NA	NA	5	3	NA	3	4	NA	1	3
P05	NA	NA	NA	NA	5	5	NA	3	3	5	5	NA	5	NA	NA	NA	NA	3	5
P07	3	3	NA	NA	3	1	1	NA	NA	5	NA	NA	NA	NA	3	5	NA	3	5
P08	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	NA	3	5
P09	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	1	5	NA	3	2	NA	NA	NA
P10	2	2	3	4	3	3	2	2	NA	5	5	3	5	4	3	3	5	3	5
P11	1	1	NA	NA	NA	NA	NA	2	NA	2	2	NA	NA	2	NA	NA	NA	1	NA
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	5	NA	NA	4	NA	NA	NA	NA	NA
P13	1	1	2	5	NA	NA	3	1	NA	NA	NA	NA	5	NA	3	5	2	5	5
P14	2	2	3	5	4	3	1	3	3	4	5	4	3	5	3	5	NA	3	3
P15	3	3	5	4	NA	NA	NA	1	NA	4	4	3	NA	5	NA	NA	NA	NA	NA

Table 5.3: Reshaped data for the descriptor *larynx height*. The ratings were normalised to the range between 1 and 5 and can have non-integer values.

	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
P01	4	2	3	3	5	5	5	3	NA	5	5	NA	NA	NA	2	2	NA	NA	NA
P02	NA	NA	2	3	4	3	3	3	NA	5	5	3	NA	5	NA	NA	NA	4	4.5
P04	NA	NA	NA	NA	5	3	2	NA	NA	NA	NA	4	4	NA	2	4	NA	2.5	4
P05	NA	NA	NA	NA	4	4	NA	3	3	5	5	NA	3	NA	NA	NA	NA	4	5
P07	3.5	4.5	NA	NA	5	3	4	NA	NA	4	5	NA	NA	NA	2	5	NA	4	4.5
P08	NA	NA	NA	NA	5	4	NA	NA	NA	NA	NA	NA	3	NA	NA	NA	NA	3	3.5
P09	NA	NA	NA	NA	NA	NA	NA	3	NA	NA	NA	2	4	NA	2	2	NA	NA	NA
P10	3	3.5	3	3	4	4	3	3	NA	5	5	3	4	5	2	3	3	3	3
P11	4	4	NA	NA	NA	NA	NA	2	NA	5	5	NA	NA	4.5	NA	NA	NA	3	NA
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	4	NA	NA	5	NA	NA	NA	NA	NA
P13	3	4	3	4	NA	NA	3	4	NA	NA	NA	NA	3	NA	2.5	3.5	4	3.5	3.5
P14	3	3.5	3	5	4	4	3	3	3	4	5	4	3.5	5	2	2.5	NA	3.5	4
P15	3	4	1	3	NA	NA	NA	3	NA	4	5	3.5	NA	4.5	NA	NA	NA	NA	NA

values (e.g. larynx height, see Table 5.3). Nominal descriptors would only have values 0 and 5. Vocal folds vibration mode was an exception here: because a voice being in falsetto is semantically closer to thin folds than to thick, using 0 for falsetto would skew the results contrary to this fact (see discussion above in this Section). Therefore, the actual values of of thick/thin vocal folds were mapped on the $[1-5]$ interval and the value for falsetto was chosen to be 9 (Table 5.4).

5.3 Inter-rater agreement: Krippendorff's alpha

For the analysis of inter-participant agreement we chose the Krippendorff's alpha coefficient. This statistical measure of agreement has been widely used in many disciplines for over 40 years. Compared to other measures of agreement Krippendorff's

Table 5.4: Reshaped data for the descriptor vocal folds thick to thin. The original 9-point scale on which the thickness was rated is normalised to the range between 1 and 5; snippets in falsetto, where no thickness was rated, are assigned the value 9.

	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
P01	2	9	3	3	5	4	1.5	4	NA	5	5	NA	NA	NA	3	3	NA	NA	NA
P02	NA	NA	1	1	3	1	3	3	NA	4	5	1	NA	3	NA	NA	NA	3	3.5
P04	NA	NA	NA	NA	4.5	2.5	2	NA	NA	NA	NA	4	4	NA	1.5	3.5	NA	2	3
P05	NA	NA	NA	NA	3	3	NA	5	2	5	5	NA	1	NA	NA	NA	NA	1	2
P07	4	5	NA	NA	9	9	5	NA	NA	3	NA	NA	NA	NA	2	4	NA	2	4
P08	NA	NA	NA	NA	NA	2	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	1	1
P09	NA	NA	NA	NA	NA	NA	NA	9	NA	NA	NA	4.5	1	NA	1.5	2	NA	NA	NA
P10	4	4	2	2	4	4	4	5	NA	4	4	2	1	2	2	3	1	2	1
P11	1	4	NA	NA	NA	NA	NA	4	NA	3	3	NA	NA	1.5	NA	NA	NA	2	NA
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	4	NA	NA	1	NA	NA	NA	NA	NA
P13	1	5	3	4	NA	NA	3	4	NA	NA	NA	NA	2	NA	2	4	4	1	1
P14	1	3	1	1	1	1	1	3	1	1	1	3	1	1	1	3	NA	1	1
P15	1	2	1.5	2.5	NA	NA	NA	9	NA	3	4	1	NA	3	NA	NA	NA	NA	NA

alpha has a number of advantages:

- it can incorporate data of different metrics (distance function)
- it can deal with missing data
- it is a generalisation of a number of standard correlation measures.

As Table 5.1 shows, some of our descriptors are described by nominal metrics while others, where Likert scale was used, adhere to interval metrics. With Krippendorff's alpha, we can use the same method to analyse both groups. Also, only one of our experts managed to rate all the tracks during the time set up for the interview, all others rated only a subset of the tracks; thus our ratings are incomplete and cannot be analysed by the vast majority of well-known statistics.

Krippendorff's alpha is defined as

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is observed disagreement and D_e is expected disagreement. Observed disagreement is the average of unit disagreements D_u

$$D_o = \sum_{u=1}^{N} \frac{m_u}{n} D_u$$

where m_u is the number of ratings in unit (data column) u and n is the overall number of pairable values. In our case units are musical snippets. Due to missing data, some cells in the data matrix will be empty (e.g. 'NA' values in Table 5.2),

therefore the number of ratings m_u for each snippet differs. Some ratings cannot be used for agreement calculation, namely those, where a single rating for the unit is available (so there are no other ratings to agree or disagree with); n is the total number of pairable ratings, for all units with $m_u \geq 2$. In our example (Table 5.2) $m_1 = 7$ and $m_9 = 2$. Since each column contains at least two ratings, all ratings are pairable and $n = \sum m_u = 129$.

A unit disagreement is the average difference δ^2_{ck} between two ratings c and k over all $m_u(m_u - 1)$ pairs of ratings within unit u:

$$D_u = \frac{1}{m_u(m_u - 1)} \sum_{i=1,j=1}^{m} \delta^2_{c_{iu}k_{ju}}$$

While observed disagreement is first calculated within units and then averaged, expected disagreement is the average difference over all pairs of pairable ratings across all units:

$$D_e = \frac{1}{n(n-1)} \sum_{u=1,w=1}^{N} \sum_{i=1,j=1}^{m} \delta^2_{c_{iu}k_{jw}}$$

with N the number of units (in our case snippets) and m the number of participants.

Usually the actual calculation of observed and expected disagreement is done by means of computationally more efficient coincidence matrix. This is a square, symmetrical $v * v$ matrix, where v is the number of possible values for the ratings of a given variable, or the number of levels of the variable. The matrix of observed coincidences contains frequencies:

$$o_{ck} = \sum_{u=1}^{N} \frac{A^u_{ck}}{m_u - 1} = o_{kc}$$

where A^u_{ck} is the number of ordered (c, k) pairs in the unit u. The coefficients of the expected coincidence matrix are as follows:

$$e_{ck} = \frac{A_{ck}}{n - 1} = e_{kc}$$

with A_{ck} the number of ordered (c, k) pairs across all pairable ratings.

Let us perform an example calculation of the coincidence matrix for our data for *subglottal pressure* in Table 5.2. The rating 1 occurs at least twice in the data columns of the following snippets: 14_1, 14_2, 18_1, 18_2, 20_1, 22_1, 34_1. Therefore we calculate:

	V1	V2	V3	V4	V5
1	7.666667	3.791667	9.596429	0.652381	2.292857
2	3.791667	1.416667	4.050000	1.084524	3.657143
3	9.596429	4.050000	17.966667	1.770238	8.616667
4	0.652381	1.084524	1.770238	1.550000	5.942857
5	2.292857	3.657143	8.616667	5.942857	17.490476

(a)

	V1	V2	V3	V4	V5
1	4.3125	2.625000	7.875000	2.062500	7.125000
2	2.6250	1.421875	4.593750	1.203125	4.156250
3	7.8750	4.593750	13.453125	3.609375	12.468750
4	2.0625	1.203125	3.609375	0.859375	3.265625
5	7.1250	4.156250	12.468750	3.265625	10.984375

(b)

Figure 5.3.1: a) coincidence matrix and b) expected coincidence matrix for our data for the physiological descriptor *subglottal pressure* (see Table 5.2).

$$o_{11} = \{unit14_1\}\frac{2}{7-1} + \{unit14_2\}\frac{6}{7-1} + \{unit18_1\}\frac{2}{8-1} +$$

$$+ \{unit18_2\}\frac{12}{8-1} + \{unit20_1\}\frac{20}{7-1} + \{unit22_1\}\frac{6}{9-1} +$$

$$+ \{unit34_1\}\frac{2}{9-1} = 7.67$$

Similarly, for a non-diagonal element we get:

$$o_{14} = \{unit18_1\}\frac{2}{8-1} + \{unit26_1\}\frac{1}{6-1} + \{unit32_2\}\frac{1}{7-1} = 0.65$$

The coincidence matrix for our data is given in Figure 5.3.1.

Employing the coincidence matrix coefficients, the Krippendorff's alpha is then given by

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_{c=1,k=1}^{v} o_{ck}\delta_{ck}^2}{\sum_{c=1,k=1}^{v} e_{ck}\delta_{ck}^2}$$

The difference function reflects the metric properties of the variable. We use two different metrics, nominal for "yes/no" ratings

$$(nominal)\delta_{ck}^2 = \begin{cases} 0, if\ c{=}k \\ 1, if\ c{\neq}k \end{cases}$$

and interval for the Likert scale:

$$(interval)\delta_{ck}^2(c-k)^2$$

For the ratings of *subglottal pressure* we used the interval metric and the Krippendorff's alpha amounted to 0.38.

For nominal metrics, Krippendorff's alpha approximates the well-known Fleiss'

kappa statistic when the number of variables (snippets) is large. For smaller sample sizes, like in our case, Krippendorff's alpha is a less conservative measure of agreement than the kappa. The same is true for interval metrics – here Pearson's intraclass correlation coefficient is more conservative than alpha for small sample sizes and is approximated by it for large sizes. In contrast to both Fleiss' kappa and Pearson's intraclass correlation, Krippendorff's alpha can deal with incomplete data, which is necessary in our case.

5.4 Statistical significance, bootstrapping, implementation

Since Krippendorff's alpha is a statistical estimate, it is essential to measure its reliability. We do it via bootstrapping: resampling the observed data. Our participants were all drawn from the pool of vocal physiology experts and performed their ratings independently from each other. This amounts to the requirement of variables being independent and identically distributed to be fulfilled. Therefore, we can use simple random sampling with replacement from the original data. Some caution has to be taken about the entities of (re)sampling though – it is not the snippets, but pairs of ratings (Krippendorff's pairable values) that have to be resampled.

We calculate the confidence interval by resampling our data 20 000 times, computing alpha for each sample and then selecting the subinterval of the alpha values distribution in which more than 2.5% and less than 97.5% of values fall. This interval serves us as an estimator of the location of the real alpha (which could only be found by interviewing all vocal experts). If it is sufficiently removed from zero and its proximity then - given that our group of singing experts is representative - an overall consensus on the given vocal production parameter can be assumed.

An algorithm for correct resampling of pairable values to calculate the confidence interval for alpha was formulated by Krippendorff and implemented by Andrew Hayes for a proprietary statistical interface SPSS. We employed a very well established open source interface called R for our statistical evaluation. R is widely used in social sciences and has overtaken SPSS as the system of choice for statistical analysis. It is supported by a large base of experienced users who contribute to further development and implement new algorithms. We found two relevant R packages: kripp.alpha that computes Krippendorff's alpha, and kripp.boot() for alpha confidence interval. Unfortunately, the latter package only implemented a straightforward

sampling of variables (in our case snippets), which, though a standard technique, is not well suited for bootstrapping of Krippendorff's alpha. We re-implemented the bootstrapping routine based on Krippendorff's algorithm and Hayes's SPSS code and contributed our implementation to the kripp.boot package for all future users[2]. Moreover, we introduced a new feature – weighted observations (see Section 5.7 for more details).

5.5 Calculating inter-participant agreement

Before calculating the inter-participant agreement, we investigated the data in form of tables and graphs for inconsistencies and contrary trends. Ratings for the velum height displayed such a trend (Figure 5.5.1). It seems that participant P14 rated velum height in the opposite way compared to most other participants. This might have resulted from the association of velum height with nasality, the former being a physiological description and the latter perceptual one: the lower the velum, the higher the nasality. It might be assumed that the participant was rating nasality instead of velum height.

We now calculated agreement and confidence intervals for each of 17 descriptors (physiological dimensions) using our R bootstrapping routine. As you can see in Figure 5.5.2 the results are not very promising. For all but two descriptors the value of alpha does not exceed 0.4, indicating that our experts did not agree about the ratings for these parameters. Also, the confidence intervals are not well separated from zero, demonstrating that these alpha values cannot be generalised beyond our experiment.

There are only three descriptors for which their confidence intervals are well separated from zero: *larynx height*, *AES* (size of vocal tract) and, somewhat borderline, *subglottal pressure*, indicating the reliability of the measurement procedure for these descriptors. With mean $\alpha = 0.44$ *subglottal pressure* has to be discarded as not displaying agreement. For *larynx height* and *AES* with the mean values of alpha being 0.59 a tendency to agreement could potentially be assumed[3]. For all other descriptors no conclusions about any kind of agreement can be made (Figure 5.5.2).

To cross-check the low agreement results we calculated alpha and confidence intervals per musical fragment (snippet), as opposed to the previous "per descriptor"

[2] `https://github.com/MikeGruz/kripp.boot`

[3] Krippendorff recommends $\alpha = 0.8$ or higher to conclude agreement, with values above 0.67 indicating a tendency to agreement

	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
P01	1	1	1	1	5	5	1	1	NA	3	3	NA	NA	NA	1	1	NA	NA	NA
P02	NA	NA	3	3	1	1	3	3	NA	3	3	1	NA	1	NA	NA	NA	1	1
P04	NA	NA	NA	NA	4	3	3	NA	NA	NA	NA	3	3	NA	1	1	NA	1	2
P05	NA	NA	NA	NA	1	1	NA	3	3	5	5	NA	3	NA	NA	NA	NA	1	3
P07	3	3	NA	NA	5	5	5	NA	NA	5	5	NA	NA	NA	1	1	NA	3	3
P08	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	NA	3	3
P09	NA	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	5	3	NA	3	3	NA	NA	NA
P10	1	1	1	1	3	3	1	1	NA	1	1	3	1	1	1	1	1	1	1
P11	5	5	NA	NA	NA	NA	NA	1	NA	3	3	NA	NA	3	NA	NA	NA	3	NA
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	5	NA	NA	3	NA	NA	NA	NA	NA
P13	5	5	3	2	NA	NA	3	3	NA	NA	NA	NA	3	NA	2	3	3	1	3
P14	5	5	5	5	3	5	5	5	5	3	3	5	3	5	5	5	NA	5	5
P15	1	1	1	1	NA	NA	NA	1	NA	4	3	2	NA	2	NA	NA	NA	NA	NA

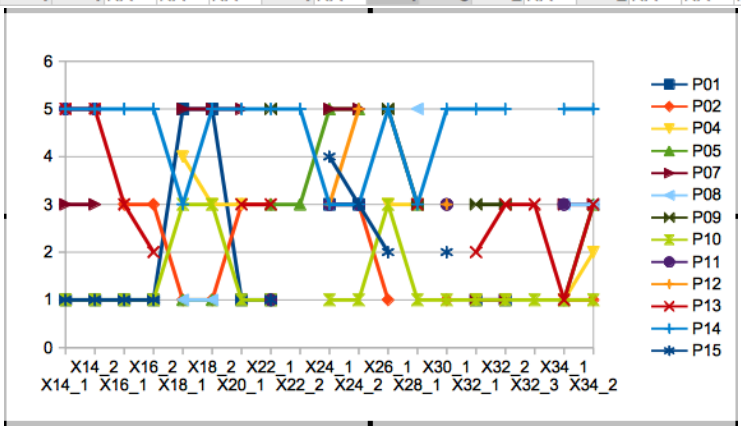


Figure 5.5.1: Velum height – participant P14's ratings seem to be contrary to the ratings of other participants ratings. It seems he was assessing nasality instead of velum height.

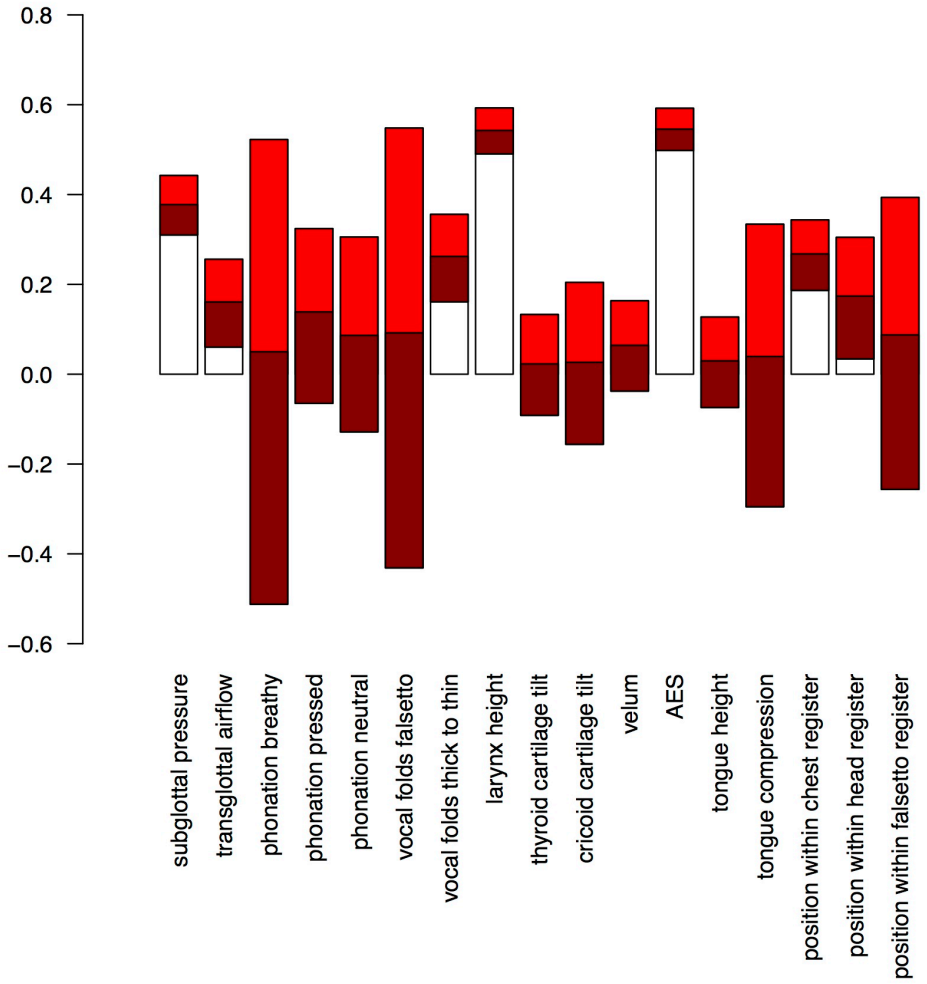


Figure 5.5.2: Inter-participant agreement for physiological descriptors. The red bar represents the confidence interval for Krippendorff's alpha coefficient for the given descriptor. These intervals were obtained via bootstrapping, with 2.5% probability lower bound and 97.5% upper bound. The line separating the dark red and the light red parts of the bar shows the value of the Krippendorff's alpha, calculated for the descriptor. The white bars under the red ones show the distance to zero. If zero falls within the confidence interval, there is a probability that the ratings are statistically unrelated. Only if the value of alpha is sufficiently large (e.g. 0.6 and above) a tendency to agreement can be claimed. In our case only two descriptors display this behaviour: *larynx height* and *AES/size of vocal tract*. The confidence intervals for these two descriptors are well separated from zero (the white bar is several times longer than the red bar), indicating the reliability of measurement, which can possibly be generalised. For all other descriptors agreement cannot be assumed.

calculation. The results (Figure 5.5.3) were considerably better than for the descriptors. If we turn to the three-dimensional view of our ratings data, we'll notice that the descriptor tables and the snippet tables are different views of the same data, the same columns in a different order. Thus, along with the low agreement for the descriptors, we were left with the puzzle, why a permutation of the same data produced such a significantly different agreement result.

First we wanted to exclude that one or two outlier snippets would cause a large alpha value drop for descriptors. We removed worst performing tracks from descriptor tables, but that didn't significantly affect the original results. Second, we looked at the differences in ratings between professional groups. We wanted to know whether there were any particular descriptors that were consistently rated differently by e.g. singing teachers and medical doctors. We introduced five participant groups and performed Krippendorff's analysis separately for each group. The results within the groups were quite similar and close to the original results for all participants.

We then sought for an explanation in the asymmetry of our data. We noticed that while descriptor tables show no particular order of present/missing data (Figure 5.2), the snippet tables had the form of filled out rows and missing rows (Figure 4.5). We decided to reshuffle the data to counter this asymmetry. In each descriptor table (participants vs. snippets, e.g. Figure 5.2) we randomly reshuffled the columns (the snippets). This permutation retains the alpha value. We put the reshuffled descriptor tables together to form a three-dimensional dataset and from it derived the new *participants vs. descriptors* tables. These tables now changed significantly, not representing ratings for a single snippet any more. Moreover, they didn't display any particular present/missing data pattern. If the mentioned data asymmetry was the reason for the difference in agreement results, we expected that the agreement for the new tables would drop, while the agreement for descriptors remained the same. Unfortunately, this was not the case. While the results changed after reshuffling, there was no clear direction in that change (Figure 5.5.4). Therefore, the present/missing data pattern in the ratings did not explain why agreement results were better per musical fragments than for physiological descriptors.

5.6 Krippendorff's alpha limitations – sparse data

Since none of the above approaches shed any light on the reason behind the low inter-participant agreement, we investigated the specifics of the Krippendorff's alpha as an estimator of agreement in general and in particular for our data. We noticed that for

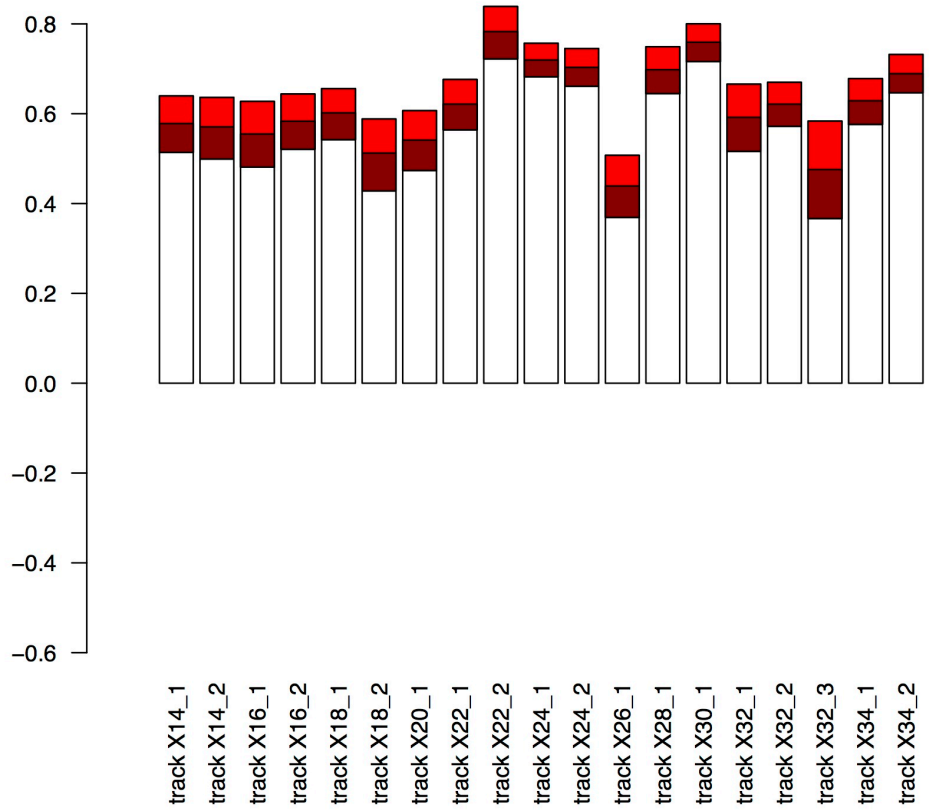


Figure 5.5.3: Inter-participant agreement for musical fragments (snippets). While the previous figure showed the agreement between participants over all tracks for each of physiological descriptors, this diagram gives the agreement between participants over all descriptors for each snippet. While we calculated Krippendorff's alpha for a permutation of the same data columns as in the previous figure, the results are better. This discrepancy needs further explanation.

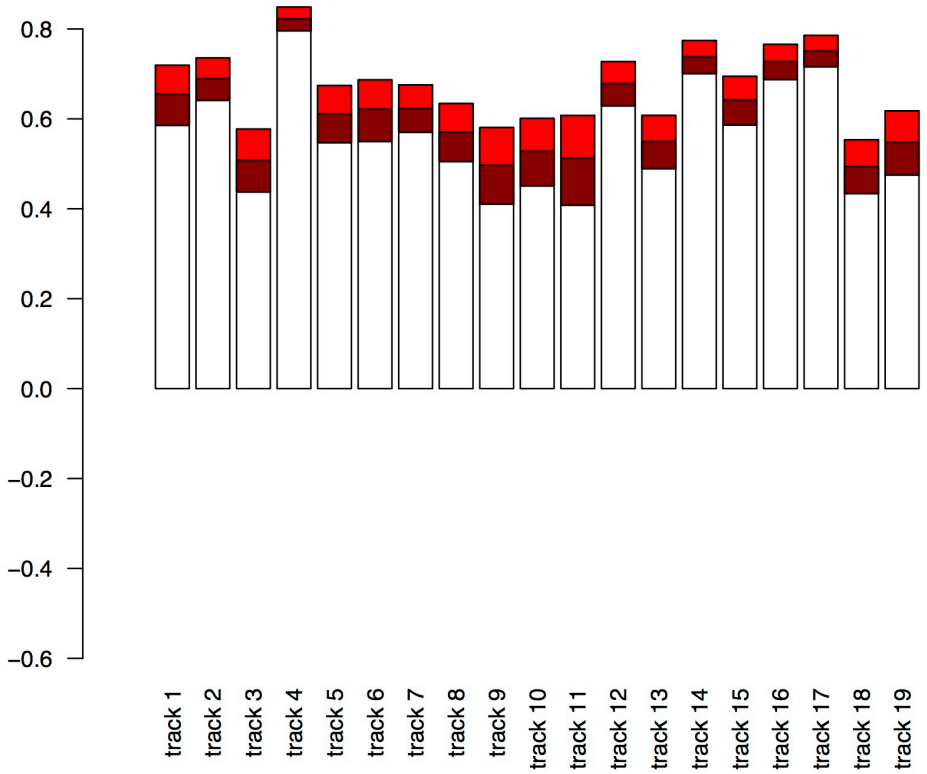


Figure 5.5.4: Inter-participant agreement for musical fragments (snippets) for reshuffled data. In search for an explanation of the discrepancy in results between the descriptor and the snippet tables, we randomly reshuffled the columns in the descriptor tables. These permutations retain the original alpha value. The new *participants*descriptors* tables do not represent the data per snippet any more and generally have no real world interpretation. But these tables do not display the same data asymmetry which was present in the original *participants*descriptors* tables. If this data asymmetry was the reason why the original per snippet data produced a better result, then we would expect the results for the reshuffled data to drop. As we can see from this figure, this expected behaviour is not present, the results here are still better than those in Figure 5.5.2.

Table 5.5: Sparse data - a nominal descriptor *breathy phonation*. The ratings are highlighted in blue and their value is indicated by the background colour. Intuitively one would think that participants agree quite well that there is no breathy phonation in these snippets. In fact for some snippets there is perfect agreement. Krippendorff's alpha though measures how randomly the values are scattered across columns: if the few positive values tended to be in one or two columns, alpha would have been high; while here the positive values are distributed more randomly across a larger number of columns, alpha is low. This example demonstrates Krippendorff's alpha's bias given sparse data.

	14_1	14_2	16_1	16_2	18_1	18_2	20_1	22_1	22_2	24_1	24_2	26_1	28_1	30_1	32_1	32_2	32_3	34_1	34_2
P01	0	5	NA	NA	0	0	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA
P02	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
P04	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
P05	NA	NA	NA	NA	0	0	NA	0	0	0	0	NA	0	NA	NA	NA	NA	0	0
P07	0	0	NA	NA	0	5	5	NA	NA	0	0	NA	NA	NA	0	0	NA	0	0
P08	NA	NA	NA	NA	5	5	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	0	0
P09	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	0	0	NA	0	0	NA	NA	NA
P10	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0
P11	0	0	NA	NA	NA	NA	NA	0	NA	0	0	NA	NA	0	NA	NA	NA	0	NA
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	NA	NA	0	NA	NA	NA	NA	NA
P13	0	5	0	0	NA	NA	0	5	NA	NA	NA	NA	0	NA	0	0	0	0	0
P14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0
P15	0	0	0	0	NA	NA	NA	0	NA	0	0	0	NA	0	NA	NA	NA	NA	NA

sparse data alpha values contradicted to the intuitive (proportional) understanding of agreement. A good example of sparse data are most of our nominal descriptors, see for instance Table 5.5 for the ratings of breathy phonation. If all participants rated a variable to be zero, one would consider it a perfect agreement. If for a given descriptor most ratings are zero, with only a small number of ratings being non-zero, one would intuitively conclude that most participants rated similarly most of the time. Alpha is calculated differently though: it is a measure of how randomly the non-zero values are distributed across the data columns. If non-zero elements tend to fall within the same columns, alpha is high; if they are scattered around a large number of columns, alpha is low. Thus, if the data contains only a small number of non-zero elements (so most ratings agree), but these happen to be scattered randomly across the table, the value of alpha drops. This contradiction disappears in situations where the data is less sparse and there are several levels of measurement.

To counter this contradiction we looked for ways to integrate the information from our sparsely rated descriptors within less sparse data. For instance, the information about the presence or absence of falsetto is represented by a nominal descriptor and is sparse, because there is not much falsetto in our singing examples. At the same time this information is also contained in the descriptor *vocal folds thick to thin*,

because if a vocalisation is in falsetto, it is assigned the value 9 (see Section 5.2 for more details).

We further introduced compound descriptors, summarising ratings for several sparse descriptors into one non-sparse descriptor. We picked out two groups of related descriptors: phonation modes and position within register. There are three phonation modes that were rated by most participants: breathy, neutral and pressed. We encoded them as three nominal variables to allow raters to assign more than one phonation mode to a snippet. There is a linear relationship between the three modes: breathy phonation corresponds to vocal hypofunction, neutral phonation to a regular vocal function, and pressed phonation to a hyperfunction (this linear relationship was not applicable to phonation modes as long as flow phonation was present). In our data we had no examples of a snippet containing breathy and pressed phonation at the same time. Therefore we decided to code the phonation mode ratings on a 5-point Likert scale with low values reflecting hypofunction and high values reflecting hyperfunction. In practice, 1 corresponded to breathy phonation only, 2 to breathy and neutral phonation, 3 to neutral only, 4 to neutral and pressed and 5 to pressed.

Position within register descriptors reflected a more complex relationship between pitch, register and vocal folds vibration mode. Similar to phonation modes, it was coded as three separate variables to allow multiple ratings for the same snippet. Though these variables had an interval metric, two of them were sparse, since much of vocalisation in our examples was in chest register, with only few examples of head and falsetto registers. To construct a compound descriptor we used the fact that these descriptors were originally suggested to capture how high a singer vocalises within a given register. The assumption was that, if a singer employs pitches which are at the higher end of their range for a given register, the singing would sound more tense and narrow, compared to the same singer in a lower range. To verify this assumption only information on the height within a register was relevant, but not the register itself. To capture the height, we took a normalised sum of the three nominal descriptors (Table 5.6).

As expected, both compound descriptors performed better than the original nominal ones. The confidence interval for phonation mode, though above zero, still lies in a close proximity to it, so no agreement claims can be made. Position within individual range, in contrast, produced a confidence interval well above zero, thus indicating a tendency to agreement between raters. The good performance of this compound descriptor was expected, since information on the register, which could

Table 5.6: Ontology adjusted, sparse descriptors replaced with compound

descriptors	**physiological dimensions**	**range**	**scale**	**metric**
subglottal pressure	subglottal pressure	low to high	5-point	interval
transglottal airflow	transglottal airflow	low to high	5-point	interval
phonation	phonation	breathy/ neutral/ pressed	5-point	interval
vocal folds vibration mode	vocal folds vibration mode thick to thin	thick/ mixed thicker/ mixed/ mixed thinner/ thin/ falsetto	17-point	interval
larynx height	larynx height	low to high	9-point	interval
thyroid cartilage tilt	thyroid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
cricoid cartilage tilt	cricoid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
velum	velum	low to high	5-point	interval
aryepiglottic sphincter	aryepiglottic sphincter	wide to narrow	5-point	interval
t**ongue**	tongue height	low to high	5-point	interval
pos. within register range	position within register	low to high	5-point	interval

cause disagreement in the original descriptors, was dropped. While an acceptable statistical approximation, the explanatory value of this compound descriptor without the register/vocal folds vibration mode information can be questioned.

Results including compound descriptors (Figure 5.6.1) instead of original nominal ones demonstrate Krippendorff's alpha bias for non-sparse data – these confidence intervals, though still far from good agreement, do not differ so significantly from the agreement per snippet results (as in Figure 5.5.3).

Figure 5.6.2 illustrates results including the new compound descriptors for four participant classes: medical professionals (otolaryngologists and surgeons, speech and language pathologists), singing teachers, Estill influenced, Sundberg influenced. The groups display distributions similar to the overall picture.

Having convinced ourselves that our statistical calculations were correct, we are still left with the puzzle how the low agreement between experts can be explained.

5.7 Confidence values

Alongside participants' ratings we also collected data on their confidence about their ratings, because in many cases physiological analysis was ambiguous for various reasons (See Figure 4.6.1). Participants rated their confidence on a 5-point Likert scale (Figure 4.5). While all our previous analysis was based purely on the ratings, at this stage we wanted to include confidence values and see whether they would have any impact on the results. There was no standard way of calculating a weighted Krippendorff's alpha coefficient, therefore we used a tailored approach. First we normalised the confidence values via a linear transformation to the [0,1] interval, with 5 mapped to 1 and 1 mapped to 0. Then we represented our data as pairs of values – these were implemented as complex numbers in R – with the real part carrying the rating and the imaginary part being the confidence value.

We adjusted our R routines for calculating Krippendorff's alpha and for the bootstrapping to be able to handle complex numbers and introduced a new metric we called *confidence*. We only collected confidence values for a subset of descriptors (including the collated *phonation* descriptor), all of which had interval metrics. Therefore, delta (distance measure) calculation for the new confidence metric was based on the interval metric and assumed complex data; but while the original interval metric was difference squared, for confidence distance measure it was multiplied with each of the confidence values:

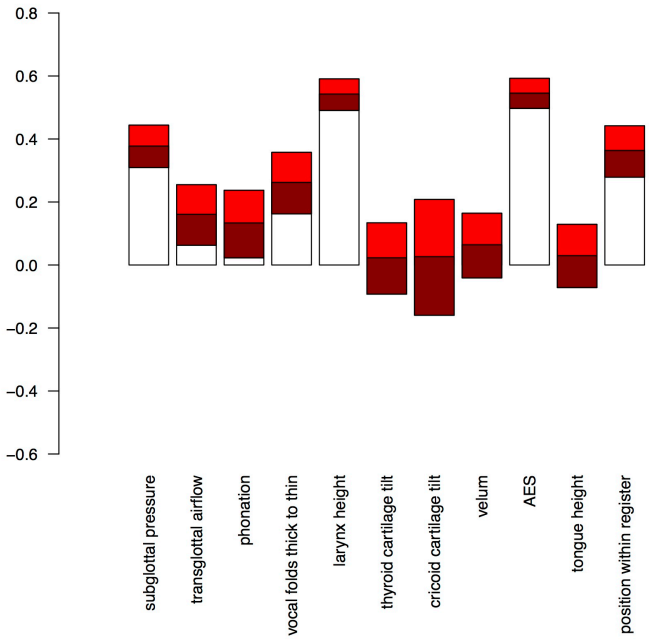
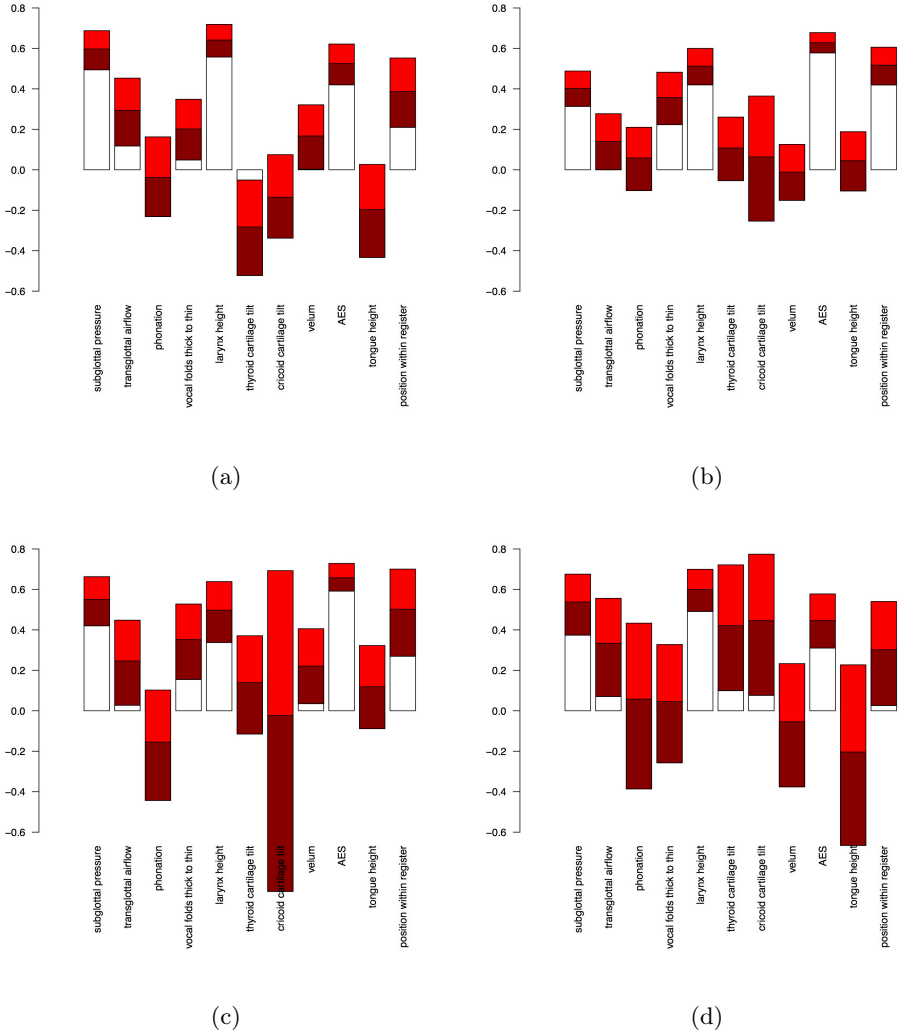


Figure 5.6.1: Inter-participant agreement with sparse descriptors removed and replaced with compound descriptors. To counter Krippendorff's alpha's bias in case of sparse data we introduced compound descriptors, where several sparse descriptors are integrated into a new descriptor, which has more levels of measurement and a less sparse representation. The result displayed in this figure does not differ so significantly from the per snippet results in Figure 5.5.3, as was the case with the original per descriptor results (Figure 5.5.2), which included a number of sparse descriptors. This demonstrates that the sparse data bias of Krippendorff's alpha was a probable reason for the above discrepancy in results.



(a) (b)

(c) (d)

Figure 5.6.2: Inter-participant agreement for four participant classes: a) medical professionals, b) singing teachers, c) Estill influenced, d) Sundberg influenced.

$$\delta_{conf}(r_1 + ic_1, r_2 + ic_2) = (r_1 - r_2)^2 * c_1 * c_2$$

The idea behind this was that we wanted the ratings to become fuzzy, the lower the confidence, the fuzzier the rating value: so that if the participant was completely convinced in his rating, the distance would remain the same, and if the participant provided a random rating value, the distance would become zero.

The results (Figure 5.7.1) did not bring any surprises: as expected the overall result has become better; the descriptors that displayed better agreement increased their agreement values; the ones where agreement values were low remained low. For larynx height and AES the mean alpha value rose from 0.59 to 0.65 and 0.64 respectively. The overall picture of the alpha distribution remained the same (see Figure 5.7.2).

5.8 Collecting reliable annotations for Cantometrics recordings

In Section 4.7 of the methodology chapter we suggested ways to proceed both in case when inter-rater agreement were good for most descriptors and when it were not. Yet our result is mixed – with two descriptors displaying good agreement and eight showing no tendency to agreement. In the next chapters we shall follow the path suggested for the case of absence of agreement, which is prevalent in our results. But let us consider the descriptors which were agreed about here.

It has been demonstrated in previous sections that two descriptors – *larynx height* and *AES* displayed a good inter-rater agreement. Our experts' annotations on these descriptors can therefore be deemed reliable given the current state of knowledge. A large musical, vocal and cultural variability in our musical examples, as well as a good diversification of raters make our argument about the reliability of their ratings even more credible. It is important to remember that these ratings do not represent any kind of truth. There is no way to measure or observe the physical process of vocal production in our musical examples; no way to determine whether our experts' ratings agree with that physical reality. More knowledge will be gained about vocal production in the future and experts' views can change. Yet at this point in time and state of knowledge, 13 experts from different countries and professions have agreed on both *larynx height* and *AES* on 19 highly varied recordings of singing. This is a good indicator that experts will continue to agree on these two descriptors

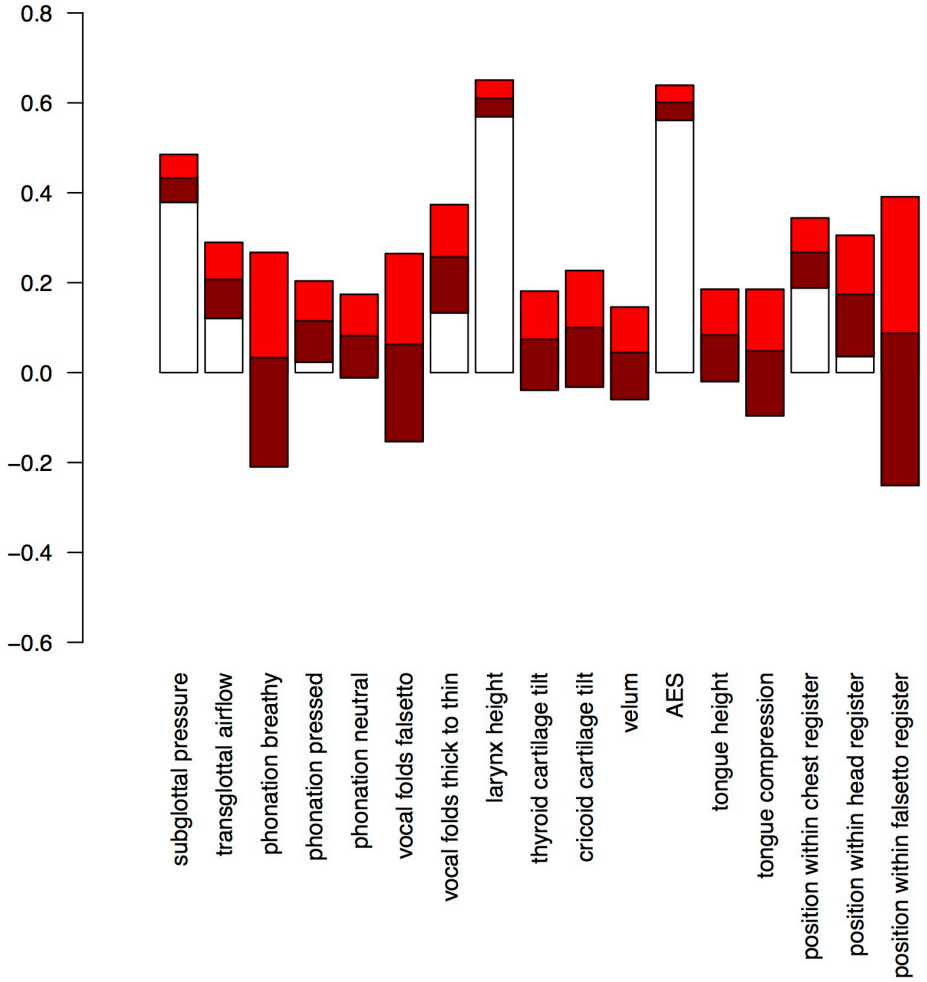


Figure 5.7.1: Inter-participant agreement taking into account participants' confidence in their ratings. Confidence values were collected for all interval descriptors with the exception of *position within register range*. They were also obtained for phonation in general, which corresponds to our compound phonation descriptor.

The results improved somewhat compared to those not taking confidence values into account (Figure 5.6.1). This improvement was expected, since the fuzziness of the ratings represented by the confidence lowered the distance between ratings. At the same time, the overall picture did not change significantly.

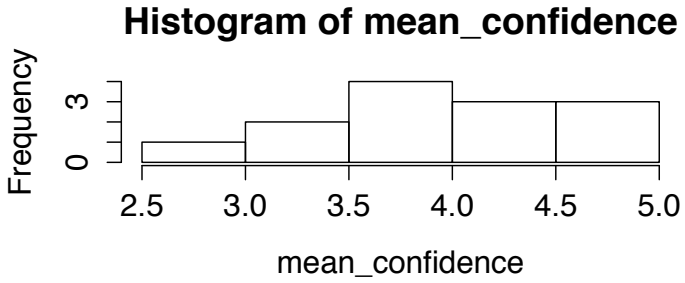


Figure 5.7.2: Mean confidence distribution for the descriptor *subglottal pressure*. The overall confidence mean for this descriptor was 3.96. Participants rated their confidence on a 5-point Likert scale, with 5 being absolutely confident and 1 meaning a random answer. This distribution demonstrates that nearly all participants had healthy doubts about their ability in the given circumstances to produce unambiguous annotations. Thus, the clearer tendency to agreement for this descriptor did not result from one or two experts providing random ratings.

if presented with other stimuli. It is also a good argument that the ratings they produced for our musical examples are the best possible estimates.

In Table 5.7 we give average ratings of *larynx height* and *AES* provided by our experts. Together with the 19 snippets these ratings constitute the first (cross-cultural) dataset with reliable annotations on vocal production. It has been published as a curated dataset at `https://osf.io/8zp7e/`.

Because there are only two descriptors with good agreement there is no need to do dimensionality reduction as suggested in Section 4.7. In order to test whether *larynx height* and *AES* ratings correlate with Cantometrics *vocal width* ratings the former have to be aggregated to a single value for the tracks containing two snippets. We performed this aggregation in accordance with the percentages attributed to each snippet in Table 4.3.2, see Tables 5.8 and 5.9. Exception was Track 22 where some participants provided ratings for the lower part – we took these ratings in account in proportion. In Track 32, two raters questioned our choice of snippet and chose to annotate a different snippet. For these two participants we took their ratings for each of the three snippets in equal proportions. As we mentioned in Section 4.3, this aggregation approach goes against Lomax's assumption that his classification would not be affected by the choice of the track segment. If he chose different segments for the Cantometrics Training Tapes, the percentage of particular

Table 5.7: Average ratings for the two descriptors that displayed good inter-rater agreement – the first reliable vocal production annotations given the current state of knowledge. All ratings were normalised to the range 1 to 5.

Snippet	Mean AES	Mean larynx height
X14_1	2.00	3.36
X14_2	1.71	3.64
X16_1	2.00	2.50
X16_2	2.67	3.50
X18_1	4.50	4.50
X18_2	3.75	3.75
X20_1	2.00	3.29
X22_1	2.11	3.00
X22_2	1.50	3.00
X24_1	4.33	4.56
X24_2	4.67	4.89
X26_1	1.67	3.25
X28_1	3.14	3.50
X30_1	4.67	4.83
X32_1	1.71	2.07
X32_2	1.71	3.14
X32_3	1.50	3.50
X34_1	2.22	3.39
X34_2	2.63	4.00

Table 5.8: *AES* ratings in comparison to Cantometrics *vocal width* classes. The tracks have been grouped and coloured to show the Cantometrics *vocal width* ratings: blue tracks were rated as wide and relaxed; orange tracks as mid; and yellow tracks as narrow and tense. Our participants' ratings were coloured in a greyscale according to the value: from 1 meaning no *AES* narrowing to 5 meaning *AES* very narrow. At the bottom are average ratings for each snippet as well as track ratings aggregated according to percentages given in Table 4.3.2.

Cantometrics rating	wide, relaxed								mid						tense, narrow				
Snippet	X16_1	X16_2	X22_1	X22_2	X26_1	X32_1	X32_2	X32_3	X14_1	X14_2	X20_1	X28_1	X34_1	X34_2	X18_1	X18_2	X24_1	X24_2	X30_1
P01	3	3	3	NA	NA	3	3	NA	3	1	3	NA	NA	NA	5	5	5	5	NA
P02	1	3	3	NA	3	NA	NA	NA	NA	NA	1	NA	3	5	5	3	5	5	3
P04	NA	NA	NA	NA	1	1	1	NA	NA	NA	1	3	1	2	3	3	NA	NA	NA
P05	NA	NA	1	1	NA	NA	NA	NA	NA	NA	NA	3	1	1	5	5	3	3	NA
P07	NA	NA	NA	NA	NA	3	3	NA	1	1	3	NA	3	3	5	5	5	5	NA
P08	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	3	3	5	3	NA	NA	NA
P09	NA	NA	3	NA	1	2	2	NA	NA	NA	NA	4	NA	NA	NA	NA	NA	NA	NA
P10	1	1	1	NA	1	1	1	1	1	1	1	3	1	1	3	3	5	5	5
P11	NA	NA	2	NA	NA	NA	NA	NA	3	3	NA	NA	3	NA	NA	NA	5	5	5
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	4	5
P13	3	3	3	NA	NA	1	1	2	1	3	3	3	3	3	NA	NA	NA	NA	NA
P14	3	5	2	2	3	1	1	NA	3	2	2	3	2	3	5	3	3	5	5
P15	1	1	1	NA	1	NA	NA	NA	2	1	NA	NA	NA	NA	NA	NA	4	5	5
Mean	2	2.6667	2.1111	1.5	1.6667	1.7143	1.7143	1.5	2	1.7143	2	3.1429	2.2222	2.625	4.5	3.75	4.3333	4.6667	4.6667
aggregated track mean	2.48		2.05		1.6667	1.694			1.9286		2	3.1429	2.3068		4.2809		4.4167		4.6667

Table 5.9: *Larynx height* ratings, see Table 5.8 for explanations.

Cantometrics rating	wide, relaxed								mid						tense, narrow				
Snippet	X16_1	X16_2	X22_1	X22_2	X26_1	X32_1	X32_2	X32_3	X14_1	X14_2	X20_1	X28_1	X34_1	X34_2	X18_1	X18_2	X24_1	X24_2	X30_1
P01	3	3	3	NA	NA	2	2	NA	4	2	5	NA	NA	NA	5	5	5	5	NA
P02	2	3	3	NA	3	NA	NA	NA	NA	NA	3	NA	4	4.5	4	3	5	5	5
P04	NA	NA	NA	NA	4	2	4	NA	NA	NA	2	4	2.5	4	5	3	NA	NA	NA
P05	NA	NA	3	3	NA	NA	NA	NA	NA	NA	NA	3	4	5	4	4	5	5	NA
P07	NA	NA	NA	NA	NA	2	5	NA	3.5	4.5	4	NA	4	4.5	5	3	4	5	NA
P08	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	3	3.5	5	4	NA	NA	NA
P09	NA	NA	3	NA	2	2	2	NA	NA	NA	NA	4	NA	NA	NA	NA	NA	NA	NA
P10	3	3	3	NA	3	2	3	3	3	3.5	3	4	3	3	4	4	5	5	5
P11	NA	NA	2	NA	NA	NA	NA	NA	4	4	NA	NA	3	NA	NA	NA	5	5	4.5
P12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	4	5
P13	3	4	4	NA	NA	2.5	3.5	4	3	4	3	3	3.5	3.5	NA	NA	NA	NA	NA
P14	3	5	3	3	4	2	2.5	NA	3	3.5	3	3.5	3.5	4	4	4	4	5	5
P15	1	3	3	NA	3.5	NA	NA	NA	3	4	NA	NA	NA	NA	NA	NA	4	5	4.5
Mean	2.5	3.5	3	3	3.25	2.0714	3.1429	3.5	3.3571	3.6429	3.2857	3.5	3.3889	4	4.5	3.75	4.5556	4.8889	4.8333
Aggregated track mean	3.22		3		3.25	2.7594			3.4286		3.2857	3.5	3.5172		4.2809		4.6389		4.8333

physiological gestures in these segments would be different. Therefore, the stability of classification in respect to the choice of the snippets should be analysed in future work.

We calculated the Spearman's correlation coefficient for the Cantometrics *vocal width* in relation to *larynx height*, *AES* and *subglottal pressure* (Figure 5.8.1). *Larynx height* displayed a strong correlation with *vocal width* ratings. *AES* also correlated well, though the values are weaker. We found no correlation with *subglottal pressure*. This is in line with our hypothesis about which descriptors contributed to the perception of *vocal width* (see Table 2.4).

In Section 1.1 of the Introduction we formulated our hypothesis – that we can

map the *vocal width* parameter to more objective descriptors of vocal production. Herewith this goal is achieved – *larynx height* is thus a good candidate to replace *vocal width*. We have shown that reliable annotations of *larynx height* can be generated by means of expert ratings. These annotations can then be used to build computational models for automatic classification of *larynx height*. This, in turn, would help to construct a large corpus of singing recordings with annotations of *larynx height*; the relationship between *larynx height* (as a substitution for *vocal width*) and the subordination of women could then be (re-)investigated.

It is important to stress the ambiguities in the definitions of both *larynx height* (see Section 2.1.3.2) and *AES* (see Section 2.1.3.2 and analysis in Section 7.2) which will have to be addressed by future work. It can also be investigated whether adding *AES* ratings to the mapping would improve *vocal width* rating prediction.

5.9 Discussion and outlook

As we have shown in the previous section, not much agreement on the ratings between the participants of our study can be claimed: of 11 original physiological descriptors (18 dimensions) only two – *larynx height* and *AES* – displayed confidence intervals well above zero. This significant lack of agreement is contrary to our expectations. In fact, we interviewed experts who had worked with vocal physiology for at least ten years (up to 45 years), and whose outcomes in their daily work rely directly on their ability to reconstruct physiological processes in the vocal apparatus. One of our basic assumptions was that these people would know what they are doing when conducting a physiological analysis of the musical examples we suggested. So why did they rate physiological descriptors so differently?

One obvious assumption of our experiment design was that physiology can be reconstructed from the auditory information in absence of other information channels – the participants analysed audio recordings of singing. Since in regular circumstances they primarily work with the singers who are present, this might have had an impact on their ability to reconstruct physiological processes. Also, the recordings they were presented with were in MP3 format and sometimes of rather poor quality, made more than half a century ago by ethnomusicologists in the field. We did not explicitly asked in the interviews whether recording quality was a hindrance, and this issue was only rarely brought up by the participants themselves.

To investigate the role of various information channels on the ability of experts to reconstruct physiological processes, a new study could be conducted which would

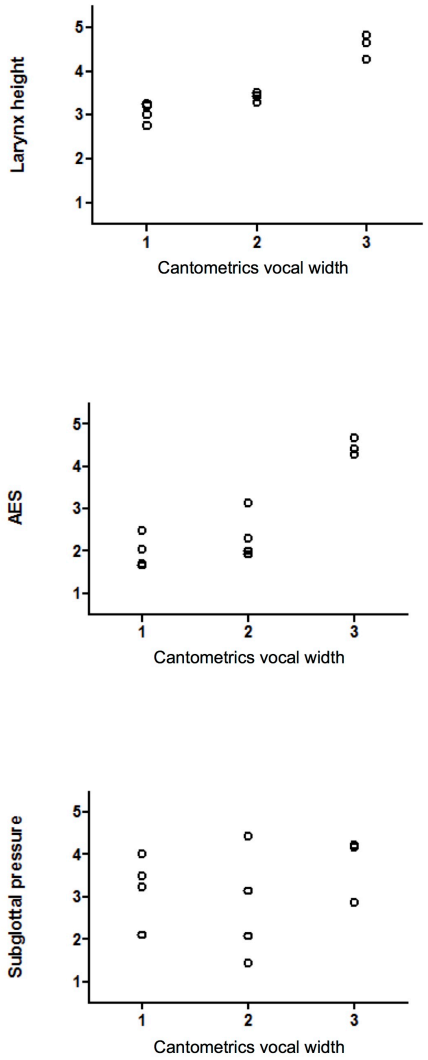


Figure 5.8.1: Correlation between reliable descriptors from our ontology and Cantometrics *vocal width*.
a) *Larynx height* vs *vocal width*: Spearman ro = 0.94, p-value < 0.0001
b) *AES* vs. *vocal width*: ro = 0.75, p-value = 0.0077
c) *subglottal pressure* vs. *vocal width*: ro = 0.17, p-value = 0.62
Larynx height displays a strong correlation; for *AES* the correlation effect is weaker; no correlation between *subglottal pressure* and *vocal width*.

include performing physiological analysis on audio recordings and on video recordings of singing. The technical setup would be similar to our study, allowing for remote interviews. At the same time, the most important information channel – the visual – would be incorporated. If the sample remains small like in our study, it is probably wise to present participants with an audio recording first, and after it has been analysed offer them a video recording of the same fragment. This approach would save time and allow for collecting more data from a small number of participants. At the same time, it would not account for possible differences/bias in the analysis as compared to analysing an unknown performance directly from a video. One could argue though that when interviewing experts such differences should be minimal. In our study each musical fragment was analysed on average by 6 participants. If this number is expected to be higher, the mentioned bias can be avoided by offering either only audio or only video for analysis in a random order. The experts can also be asked direct questions on the influence of the presence of the video information on their analysis – it might be best to include an open question into the wrap-up block.

Including analysis of live singing would be desirable, though the technical setup for this kind of experiment would be much more complex, also making it harder to control the conditions. It is important to be aware of the cultural component inevitably present in any such study: it either should involve experts and singers proficient in the same singing tradition, then claiming results for this tradition only, or, if cross-cultural encounters are desirable, they should be carefully planned, documented and interpreted.

If the influence of missing visual and other contextual information on the ability of experts to reconstruct physiological processes can be verified, it will have significant implications for a number of disciplines, including singing education and musical analysis of singing performance. We may learn whether singing lessons via Skype could work, what the differences are between teaching an individual or a choir, and possibly get insights into the attraction of the popular TV show "The Voice", where the judges are not allowed to see the contestants. But the biggest warning sign will be given to music informatics (MIR): this field relies in its research often exclusively on large corpora of audio recordings of music and singing. If experts' analysis of a small dataset of recordings provides largely random results, what can be expected of an analysis of a large dataset by regular music listeners? Any interpretations of such results must be very well grounded in arguments other than statistical to be of any explanatory value.

Returning to our main assumption – experts in vocal physiology know what they are doing when they perform physiological analysis. Whether this is true is a psychological question that won't be easy to untangle. First, we could employ participants' own assessment of their confidence, which we collected together with their ratings but haven't analysed separately yet.

Considering the problem of reconstructing physiology from the resulting acoustic phenomena from the point of view of voice science, it must be noted, that more than one solutions may be possible. For example, David Howard mentioned in personal communication, in conjunction with his real-life vocal tract area display (Howard et al. 2004), that deconstructing the form of the vocal tract from an audio input of singing by means of his mathematical model sometimes led to multiple solutions. There seems to be no clear way so far to analyse which of these solutions are physiologically possible and which are realised in a given cultural practice. Still this is a good indication that multiple ways of achieving similar acoustic results could be used by singers. Controlling for this fact experimentally represents a tricky task, requiring experienced singers who can vary and control their physiology while singing with a stroboscope inserted into their vocal tract. A way to approach this question through our data would be to investigate salience of physiological descriptors: during the interviews we asked participants which physiology was more / less significant for their perception of singing in the track being open and relaxed or being tense and narrow. They provided free-form answers listing salient physiological descriptors. These answers could possibly be reformulated numerically and analysed statistically, e.g. in relationship to participants' physiological ratings. Such analysis, if successful, could provide insights into the question of conscious choice of particular salient descriptors or descriptor configurations as opposed to a random rating.

And finally, the role of cultural bias in our study remains open – it cannot be uncovered statistically on such a small sample. Our original assumption – "the experts know what they are doing" – was based on the idea that if someone is able to reconstruct physiology from audio, they would be able to use their skill in any context, for any vocal production. This is in fact a very strong assumption that might prove inappropriate. One obvious weakness is that experts' musical experience and the area of musical practice of their clients/patients/subjects is the context in which their expertise and its scope were formed. It does not guarantee success in other musical contexts, e.g. if a participant is presented with singing from another culture, that is very different to participant's background. To verify it scientifically, another study with the same setup as ours could be performed, but only including musical

examples from Western music; Western classical and popular music was part of each participant's musical experience. The inter-participant agreement could then be compared between the two studies. In fact, we did have an example of singing in our study (track 32), which, though coming from Ukraine, followed closely Western classical canons. We also carefully documented participants' musical and cultural biographies.

We present the qualitative analysis of the interviews in Chapter 6 where a close examination of what participants said in the interviews, in particular their answers to open questions, are investigated systematically. Chapter 7 is dedicated to the meta-analysis of confounding issues and underlying reasons for disagreement. We discuss those findings in Chapter 8 from the point of view of different disciplines. But before we move on to qualitative analysis we would like to suggest an experiment to test inter-participant agreement within well-known musical genres to exclude the cross-cultural variation. This experiment is based on what we learnt from the work described in this chapter.

5.10 Future work – inter-participant agreement for well-known musical genres

The main difference between our current study and the proposed experiment is the cultural component: while for this study we deliberately chose musical examples from various cultures, of a wide geographic distribution, assuming that most participants won't be familiar with the majority of these cultures, for the new experiment the task is the opposite – to make sure that musical examples come from a culture and a genre that are part of participants' (and participants' clients/patients/subjects') musical background and are well known and understood. If inter-participant agreement for physiological ratings is well above our current results, it will demonstrate that previous knowledge of the culture from which the singing performance originated has a significant influence on the ability of experts to reconstruct physiological processes from the acoustic result.

The primary decision to be made by the researcher conducting the proposed experiment is how exactly to control for the familiarity of musical style in the examples presented to the participants. It has to be specified, how broad or narrow the pool of familiar examples should be in relation to each participant. Thinking more broadly, one can argue that all participants are of Western background and therefore fa-

miliar with Western classical and popular music. At the same time, we know of examples where teachers of classical (opera) singing struggle when presented with or required to teach contemporary commercial music. Possibly, this should be the first differentiation. Both classical and popular music consist of a large number of genres and styles. If approaching the task narrowly, we have two options. The first is to devise a list of genres for both popular and classical music; this list can be borrowed from a library catalogue/thesaurus (such as Library of Congress Subject Headings) or from an online music archive/distributor. The participants are then asked prior to the interviews to tick the boxes with the genres they are familiar with in the context of their daily work. Another option would be to ask participants in a free form to name the repertoires they are well acquainted with through their profession. Then overlaps have to be identified. If the strict categories do not display any overlaps, these categories will have to be generalised: in case of a library list, a thesaurus can be used for a controlled generalisation; if categories were generated by the participants themselves, though they are more authentic and really mean something to the participants, generalising these categories is less straightforward and may require another round of contact between the researcher and the participants. Musical examples will have to be chosen from the overlapping genres (possibly after generalising the original categories wide enough to enable overlaps). To ensure detailed control of familiarity, it is advisable to ask the participants to describe their involvement with the genres of the examples chosen for analysis.

The bottleneck of the whole procedure will be finding vocal physiology experts and, since they are all successful busy professionals, their time. We were lucky to find 15 experts who were happy to spend 90 minutes analysing our musical examples, of them 13 interviews took place. Participants' pace of analysis varied a lot, from rating two fragments to rating all 19 fragments within 90 minutes. Some offered more time to continue the interviews. On average participants in our current study rated 6.15 snippets with standard deviation 2.64. To limit the time the experts have to spend on the interviews in the future experiment (to e.g. 90 minutes), the researcher has to take these variations into account. Participants cannot be expected to rate all musical examples, therefore, like in our present study, some data in the resulting tables will be missing. To make sure that the ratings are independent of the order in which musical examples are presented some randomisation of the order will have to be introduced. At the same time each track needs to be annotated by at least two participants (ideally more) to enable the inter-rater agreement analysis. The limited time and the number of experts constrain the number of musical examples that can

Table 5.10: This table allows a preliminary estimation of the tradeoff between the number of participants and the number of tracks. The numbers are recalculated as of Table 4.1 based on the empirically calculated expectation of how many tracks can be rated during an interview. In our study experts rated on average 6.15 tracks per interview.

The table gives the values of the probability that a track is annotated by less than 5 experts in a study. It is desirable to keep this probability low when planning a study, e.g. below 5%, so that tracks are annotated sufficiently frequently for inter-participant agreement to be meaningful. At the same time a larger number of tracks allows for a better representation of the ontological space in the study. The optimal values (the highest values below 5%) are marked green.

		number of participants																
		10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
no. of musical examples	8	0.013	0.005	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	9	0.061	0.029	0.014	0.006	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10	0.142	0.082	0.046	0.025	0.013	0.007	0.003	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	11	0.242	0.158	0.100	0.061	0.037	0.021	0.012	0.007	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000
	12	0.346	0.247	0.170	0.115	0.075	0.049	0.031	0.019	0.012	0.007	0.004	0.003	0.002	0.001	0.001	0.000
	13	0.443	0.337	0.250	0.180	0.127	0.088	0.060	0.040	0.027	0.017	0.011	0.007	0.005	0.003	0.002	0.001
	14	0.529	0.424	0.330	0.252	0.188	0.138	0.099	0.071	0.050	0.034	0.024	0.016	0.011	0.007	0.005	0.003
	15	0.603	0.502	0.408	0.325	0.253	0.194	0.147	0.109	0.081	0.059	0.042	0.030	0.021	0.015	0.010	0.007
	16	0.664	0.571	0.480	0.396	0.320	0.255	0.200	0.155	0.118	0.090	0.067	0.050	0.037	0.027	0.020	0.014
	17	0.714	0.630	0.544	0.462	0.385	0.316	0.256	0.205	0.162	0.127	0.098	0.075	0.057	0.044	0.033	0.024
	18	0.754	0.680	0.601	0.522	0.446	0.376	0.312	0.257	0.209	0.168	0.134	0.106	0.083	0.065	0.050	0.039
	19	0.786	0.721	0.650	0.576	0.503	0.433	0.368	0.309	0.257	0.212	0.173	0.140	0.113	0.090	0.072	0.057
	20	0.811	0.755	0.691	0.623	0.554	0.486	0.421	0.361	0.307	0.258	0.215	0.178	0.146	0.120	0.097	0.078

be processed (see Section 4.2). In Table 5.10 we recalculated the lower bound probability (the probability that a track is annotated by less than 5 participants) using our empirical data, in particular the expectation of how many tracks are annotated per interview.

Along with an appropriate control for familiarity and diversity of musical styles, the examples should be chosen to represent the physiological space, spanning as many ontological dimensions as possible. The ontological space has to be similar to the one we suggested to retain comparability of results. Musical examples will have to be pre-processed to extract fragments for physiological analysis (Section 4.3). The practicalities of data collection can be borrowed from our current study in its entirety (Section 4.6). We recommend that the interviews are audio recorded to enable qualitative analysis, since the interpretation of the quantitative results depends strongly on the understandings that can only be gained though qualitative work. Ideally, two recording devices should be used to ensure a backup recording in case of failure. Also, using two exemplars of the physio analysis form filled out simultaneously by the interviewer and the participant during the discussion of the

ratings was a very useful tool.

After the data has been collected and recorded, the inter-participant agreement can be analysed by means of our Krippendorff's alpha bootstrapping implementation in R (Section 5.4). If the data format is the same as in our study, our R script automating the calculation for all descriptors and the plotting can be used to create an illustration like in Figure 5.5.2. Comparing the new results with this Figure will tell a story about the cultural component of the physiological analysis of singing: whether the experts can better agree about the underlying physiological processes for recordings of singing from their own culture as opposed to other singing traditions, with which they are less familiar.

6 Interviews – qualitative analysis

This chapter is concerned with the qualitative analysis of the data collected in the interviews with 13 vocal physiology experts. The interviews were conducted for our investigative study that implemented the integrated approach to revising Cantometrics and employing vocal physiology for automatic classification of singing (see Section 1.4). We investigated whether reliable annotations of vocal production can be produced for the original Cantometrics recordings. Yet the significance of the results goes beyond Cantometrics and MIR.

Chapter 4 lays out the methodology for our mixed-method study on expert knowledge elicitation for the physiology of vocal production. Quantitative analysis of the study is performed in Chapter 5: it concludes that only for two out of 11 descriptors of vocal physiology a tendency to agreement between experts could be established. The consequence of this puzzling dissonance between physiology experts is that there is no credible way at this point to determine what physiological processes took place in the recordings of singing they were presented with. This finding can be generalised to other datasets of recordings with similar variability: given the current state of knowledge on vocal physiology, expert annotations are not reliable/consistent and cannot be used for automatic classification models. It also raises questions about our ability in general to de-construct internal physiological processes auditorially-perceptually (based on listening) - something vocal experts are routinely engaged in.

It is particularly important to investigate possible reasons for disagreement. We expected our qualitative data to give us insights into these reasons. This chapter is dedicated to a close-to-the-data analysis of the interviews. The aim is to get a better, more nuanced understanding of participants' view of physiology, to identify common themes, to analyse problem cases and disagreement, to acquire an insight into the causes for it. Such a detailed analysis and deep understanding opens up new paths for further research, which will be discussed at the end of the chapter. Meta-analysis is then presented in Chapter 7.

We largely kept to the qualitative analysis plan presented in Section 4.8 of the

Table 6.1: Participants' perceptual ratings of tension and narrowness of vocalisation in Track 24: on a Likert scale, 1 meaning none and 5 meaning a lot. Participants agree about the singing being narrow, but are less unanimous about tension.

participant	P01	P02	P05	P07	P10	P11	P12	P14	P15	mean	std
tension	3	5	5	5	1	1	3	5	4	3.5	1.67
narrowness	3	5	5	5	5	5	3	5	5	4.5	0.88

Methodology chapter. That section also explains the iterative implementation of the analysis stage in our study.

6.1 Track 24 – beautiful women in Northeast Thailand

To offer a better focus and clarity, the analysis in this chapter will sometimes use Track 24 from our dataset as an example, which produced the largest number of interpretations and led to numerous discussions in the interviews.

Track 24 is a 38 seconds long fragment of a song from Northeast Thailand. The recording was taken from the Music of Thailand collection (Folkways FE4463) recorded and edited by Howard K. Kaufman in 1960. Lomax gives the following description of its context:

> "A country girl from this highly stratified, irrigation culture sings in a tense voice indicative of the sanctions and responsibilities that weigh upon S.E. Asian women. Her song thanks Buddha for the beauties of his creations—especially women. Mouth organ accompaniment." (Lomax 1977)

This description reflects Lomax's judgement of the singing on the recording being tense, and also displays his belief in his insights and findings on the situation of women in relation to singing: the singing is tense, because the culture is highly stratified and numerous sanctions and responsibilities weigh upon women in Southeast Asia. In the Cantometrics database this recording was rated as tense and/or narrow, the vocal width factor being low.

The respondents of our study, while mainly confirming the narrowness of the sound, sometimes disagree on it being tense (Table 6.1).

These perceptual ratings were collected before the physiological analysis, after listening to the whole track, therefore presumably under similar conditions as the

Cantometrics ratings. Our ratings are on a 5-point Likert scale with 1 meaning no described quality being perceived and 5 being the maximum amount of the quality. As we can see the narrowness, with the mean of 4.5 and a lower standard deviation points to the singing being consistently perceived as narrow. At the same time the ratings of tension cover all the spectrum of values and do not display a clear trend.

Those participants who rated the singing to be less tense describe it as being comfortable and efficient:

> "She's quite comfortable there, though it's very tight, a very small place, she's comfortable" (P11)
>
> "that's a very efficient sound" (P10)
>
> "High laryngeal position, quite efficient that. I don't think she stretched at all, I think it's in a high laryngeal position but I think she could go strictly higher if she had to, I suspect." (P12)
>
> "If you think about the linguistic patterns and the voice quality that someone from that culture uses, probably for them it's not very tense. It sounds tense, to us, to a Western ear that sounds tense." (P15)

And P01 continues the thought that the language influences the singing:

> "Maybe women speak on a higher pitch in Chinese? Then their habitus is already very different."

Those for whom this vocalisation is tense mention the reason for the disagreement:

> "This is absolutely squeezed and I have to say that also she sounds tense in my ears, in my culture (laughs)." (P05)

And confirm the efficiency:

> "I perceive it as tense and squeezed, but I would like to add that it completely goes against my own... it's a paradox for me because I think that if she is used to this in her culture and so on, I don't think that she is definitely being tired for example of singing in this style." (P05)

6.2 Physiological descriptors – introduction

One of the declared aims of our experiment was to verify the appropriateness and the usefulness of the ontology we suggested for rating physiological configurations of

vocal apparatus. In the following sections we shall look in detail at each descriptor of our ontology, discuss our participants' opinions on them, analyse difficulties and uncertainties in rating, address specific questions which are raised by the experts in relation to descriptors. We shall also summarise participants' suggestions to possible additions or changes to the ontology.

As mentioned in Section 5.1 our ontology was generally well received by the participants and most of the terms were familiar: over 80% of participants were happy to rate over 80% of the terms. Some have expressed their approval of our choice of terminology explicitly or implicitly, confirming that they don't see the need to make changes. In some cases reformulating the terms in less technical words (size of the vocal tract for AES, yelling in the sound for cricoid tilt) was necessary. The descriptors that participants had the most difficulties with were thyroid and cricoid cartilage tilts: some interviewees were not familiar with the terms; others expressed doubts about their feasibility.

One of the experts suggested to structure the ontology into descriptors related to the vocal source and those affecting the resonance body. We shall follow this approach here. We begin our analysis with descriptors that define the basic physiological setting: sound source pressure and airflow, phonation, vocal folds vibration mode and register. We then continue with descriptors related to the resonance body, e.g. to the brightness of the sound: position of the larynx, thyroid and cricoid cartilage tilts, AES, velum and tongue. We shall round up with participants' suggestions for expanding the ontology.

6.3 Vocal source – Pressure, Airflow, Phonation

Subglottal pressure and transglottal airflow are the primary characteristics of the vocal source in singing. They were thoroughly studied by Johan Sundberg in his seminal book "The science of the singing voice" (Sundberg 1987). The Estill model does not employ these terms, but almost all of the participants, including those working in the Estill system, not only knew but also actively used them. The only exception was the singing teacher related to Catherine Sadolin's Complete Vocal Technique (Sadolin 2000) – in this system vocal source is not considered independently, but only as part of the whole vocal apparatus. Participants' opinions on the modes of phonation in particular singing examples seemed to diverge significantly: our quantitative results show a slight tendency to agreement about the values of subglottal pressure, but no agreement was found for transglottal airflow.

6.3.1 Subglottal pressure vs transglottal airflow

For our chosen Track 24 participant P12 gives the following account of the subglottal pressure:

> "It's high-ish. There is quite a long closed phase."

And on transglottal airflow he notes:

> "On the lower side of moderate. She's quite keen to do very long phrases, so she's clearly got long form, isn't throwing it out. There is quite a long closed phase I suspect. So it can't be high."

A long closed phase means that the vocal folds are closed for the most part of the vibratory cycle. Therefore there is only little time for the air to escape through the folds. Similarly there is more time for the air to build up pressure behind the closed vocal folds. This is a typical picture of a long closed phase characterised by a higher pressure and a lower flow.

Quite often participants named pressure and glottal flow as salient descriptors.

> "The crucial thing to be able to sing this way was the low breath pressure." (P01, Track 18)

There is one important issue about the pressure and the flow on which participants had differing views. Some believe that these descriptors are directly related – the higher the pressure, the lower the flow, and vice versa. E.g. P10 says:

> "according to the laws of physics... subglottic pressure and transglottal airflow have to work in inverse proportion to one another."

There is certainly some truth in this, as we have just seen on the example of Track 24: the longer the closed phase of the vibratory cycle, the less air escapes, the more pressure builds up and vice versa. Other participants though do not think that the two parameters are inversely proportional. This is reflected in their ratings, which can have a low pressure and a low airflow in one snippet. E.g. P12 sees more parameters to the equation, in his view vocal folds thickness and length are also part of the relationship:

> "... it's not as thick as it would be in a full chest vocal bit. But it's a compromise, you're injecting more thickness and shortness into the gesture. There is a reciprocal relationship to airflow, resistance and pressure, and you can't change one without at least one of the others being changed. Yes, we learn how to play with that."

It must be noted that a large part of Sundberg's research is devoted to exactly this relationship, what he studies is the deviation from that inverse proportionality. All his notion of phonation modes is based on the discrepancies that arise from pressure and glottal flow not being directly related. This brings us to our next descriptor – phonation modes.

6.3.2 Phonation

Phonation has been a subject of study in speaking voice for a long time. Johan Sundberg was the first to apply the terms within academic research of singing voice. He introduced a new phonation mode – *flow*, with which he attempted to capture the vocal strategy of Western classical singers in comparison to other styles. Unfortunately many of our participants were not familiar with the *flow* mode and we did not gather enough data. For the analysis here we concentrate on three other modes: breathy, neutral and pressed. These seem to belong to general knowledge among our participants, even those who were not familiar with Sundberg's work were comfortable using the terms.

There was no agreement about the ratings of phonation modes whatsoever among our participants. For instance, for our chosen Track 24, 4 out of 8 raters thought that the first snippet was in neutral phonation, while other four classified it as pressed phonation. Interestingly, those who voted for neutral phonation felt that they had to comment on it, implicitly expecting others to reference it as pressed sound. This is probably due to the fact that pressed phonation is often considered as wrong, inefficient or unhealthy singing:

> "The term pressed phonation to me infers constriction, infers a reduced freedom of the vibrating portion of the vocal folds. I'm absolutely not hearing that there, that's a very efficient sound. ... I wonder how many people would label it pressed phonation, because it's unpleasant and out of tune." (P10, Track 30)

P01 agrees mainly:

> "This is not a pressed sound. It is all quite relaxed, in particular because the vocal folds are thin. There is a little moment though, where she goes up high, it feels like there is some pushing there."

This "little moment" caused the participant to rate the second snippet as pressed.

P05 also rejects to call it pressed phonation. She even doubts the usefulness of phonation mode categorisation in this context:

> "Even if I can call it pressed, it's not, I don't think it's categorising the sound quality if it's pressed or neutral. That is less important I think." (P05)

None of the experts who rated it pressed felt the need to comment. In Track 30 P10 explicitly expresses his expectation that other experts would see it as pressed:

> "I wonder how many people would label it pressed phonation, because it's unpleasant and out of tune."

He thinks it is a very efficient sound. But what he implicitly means is that people sometimes tend to label any vocal production they consider aesthetically or technically unacceptable as pressed phonation.

In Track 16 clinicians among the participants associated pressed phonation with vocal health problems:

> "almost pressed phonation, but probably caused by age-related vocal health issues like reflux or a condition of the laryngeal tissue." (P10)

P12 agrees:

> "For me, it's pressed but I would associate that with various voice problems rather than anything else, where I know nothing about the language or if you'd have seen it."

This is in line with the common use of term pressed phonation in particular among phoniators and other medical professions. Though, as we have seen, when confronted with vocal production from unfamiliar cultures, that work with levels of pressure unknown in Western music, some experts refuse to call it pressed or doubt the usefulness of this terminology.

Despite obvious disagreements about phonation modes of singing excerpts presented to them, our experts mostly expressed high confidence in their ratings. It could therefore be assumed that the differences in ratings are due to different interpretations of the term phonation. For instance, we have seen pressed phonation as a label of vocal health issues, of inefficient, unprofessional or "incorrect" singing, as well as a non-judgemental characteristic of vocal production based on vocal source physics. All these interpretations took place without any reliable measurements, based

solely on experts' auditory experience, complicated by the fact, that they were not familiar with the musical cultures from which singing examples originated. It is not surprising that these varying interpretations have lead to different quantitative results.

6.4 Vocal folds vibration mode and register

The most fundamental aspect of vocal production, and probably the most argued about, is the registration: why, how, where and whether at all the voice changes gears when a singer goes up in pitch. Classical register theory (see Section 2.1.2) talks about two (modal and falsetto), three (chest, head and mix) or even five (adding fistula and vocal fry) registers. In contrast, Jo Estill's physiological approach rejects the notion of registers. Instead she talks about the thickness of vocal folds which can be adjusted gradually, claiming that a singer can be trained to vocalise with any vocal folds thickness on any pitch (Section 2.1.3.2). While adjustments in vocal folds thickness do not constitute a change in physiological mechanism for Estill, she does mark out falsetto as a separate phenomenon, due to two physiological changes she observed: a) vocal folds only touch each other with their edges and b) the larynx is tilted thus changing the plane in which vocal folds vibrate.

We included both views in our ontology. The *vocal folds vibration mode* descriptor reflects the Estillian view that does not recognise registration. To reflect its complexity two dimensions were rated: one nominal differentiating between falsetto and a modal vibration mode; another, only applicable to the modal mode of vibration, putting the thickness of the folds on a linear scale. Our *position within register range* descriptor in turn was based on the classical knowledge of registration, differentiating between chest and head voice as well as falsetto.

6.4.1 Register or not?

We begin with an assessment of our participants' views on registration and its physiological mechanisms expressed in their interviews.

We were reluctant to use the term register given all the critique. Yet many participants confirmed that they were happy to use it, that they used it in their daily work and that it even may be one of the most salient descriptors. They were not unanimous on that though. The goal of our register descriptor was originally not at all about registration, we were interested in the range, looking to rate how

high in their individual range the singers were vocalising. The idea behind it was that the higher a singer goes in pitch, the more compelling the task of vocalising becomes, the more tense the voice tends to sound. But then, when the voice changes gears, it is reset – this is actually the reason why the *passagio* happens – the voice cannot continue in the given physiological setting. Therefore, the question of the range only makes sense in the context of register. But how do we negotiate it with the experts who do not accept the concept of registration?

In our very first interview we tried to avoid a direct reference to register while explaining the range descriptor. After a long discussion our participant concluded:

> "Is it going in the direction of chest voice, head voice? Like, say, register? ... In my opinion you can absolutely ask in what part of register the sound moves around. I still assume there are registers. And there are particular regularities for different registers. On the lower notes I feel the sound in my chest, and on the higher notes in my head, and this is not going to change" (P01)

We asked all our participants whether they would be happy to rate range based on registers and all of them were comfortable with that. Some didn't even see a big difference between the register concept and the physiological thick/thin folds approach in the context of our study:

> "Mode of vocal folds vibrations is the same as register as far as I'm concerned." (P15)

She illustrates the mapping between the two descriptors on the example of our chosen Track 24:

> "In modes of vibration I talked about moving between mixed thicker and mixed thinner, and if we're going to relate that to register mechanisms we're moving between chest and head."

Other participants constructed different interpretations of both vocal folds vibration mode and register descriptors based on their experience:

> "I have rated the latter part [the register descriptor] as if I would see a stroboscopic picture, and for the former [vocal folds vibration mode] I imagined how it could be described in one dimension." (P02)

P10 points out that there is a lot of confusion about the terminology, because different singing teachers describe registration in different ways. He also remarks that registration is often seen in relation to range, which he declines:

> "I actually hate attaching registration to pitch. I know that that is the habit for most people. But in a trained voice that chest register can go as high as the falsetto or the head or whatever. Head can go as low as the chest, even though it is not a very attractive place to be, it's possible. So I would think about it across the whole spectrum of pitch."

It was not unusual that register was named among the most salient descriptors. For instance, P11 supported our idea that the *position within register range* would affect all the physiology:

> "Probably on this list I would say the pitch range comfort, like the middle of his chest register. Because if you can tell that then you're gonna assign a lot of other things to the fact that maybe on a lower pitch range the same person would be quite comfortable, but just because he's up higher than he normally sings, all these other things are going to get skewed. If I had to choose one, I would choose the range." (P11)

6.4.2 Chest vs Head vs Falsetto

Participant P14 differentiated between head and chest voice or register. Chest register was for him the configuration of the vocal apparatus typical for speech; he also called it modal register. Head voice would only be used by trained Western opera singers in a higher tessitura:

> "This is kind of a trained Western singing. He would lower his larynx, he would have a little bit more [transglottal air]flow, and not compress his vocal folds as much, allow for the vibration; more like aerodynamic forces to work [...]." (P14)

According to P14 head voice is characterised by a tilted spectrum with a high fundamental, as opposed to non-opera singers in chest voice, where the fundamental is low and the spectrum is dominated by other overtones. The two registers can be mixed given the singer is able to produce sounds in both of them. Falsetto on the other hand is different from the head voice in that the vocal folds are barely touching each other, while in the head voice the closure is tighter. Vocal folds are

stretched long and thin in both head voice and falsetto, but in head voice a larger part of the vocal folds vibrates, allowing for more vibrato. He speaks of untrained falsetto and speech falsetto, differentiating it from head voice.

Other participants agree with the above view on falsetto:

a) that only a part of the folds' mass vibrates

> "We call this flagiolet, and that is where the vocal folds only vibrate on a certain part of it." (P07, CVT)

b) that vocal folds only touch each other at the edges

> "In falsetto the closure occurs exclusively via the edges of the folds, and the folds are placed in the way that not their entire mass vibrates. The difference to thin folds is that though they are thin, they still have a regular contact regarding the surface epithelium." (P02)

P12 describes the physiological mechanism of modal voice vs falsetto in more detail. In modal voice the closure of abducted vocal folds as well as the opening of adducted folds begins from beneath, the wave of the epithelium spreading upwards. This aerodynamic process is characterised by non-linearity, by phase reversal being a crucial part of the process.

> "If you've got a modal voice, it opens from beneath. The wave spreads up, puts them through, and even while it's breaking over the top, it's closing there and coming back. Interestingly, that means there is a phase reversal, a non-linearity, you're going into non-linear dynamics. ... That's by definition a modal voice. It doesn't matter what the pitch is, that's a modal voice because of phase reversal." (P12)

When the touching area of the vocal folds gets thinner and the wave gets shorter and shorter at some point it ceases to exist and this is the passaggio. In falsetto there is no phase reversal. The vocal folds just touch each other.

Thus the thickness of the folds also defines the touching surface during vocal folds closure: thick folds display a full closure, a large closure surface and a longer closure wave; the thinner the folds, the narrower the closure surface and the shorter the wave.

Jo Estill introduced another factor differentiating falsetto from thin folds: the plane in which the vibration takes place. In her opinion the plane is changed for

falsetto. An Estill system adept among our experts operates routinely with this aspect:

> "In the second snippet the singer changes from chest voice to falsetto, though only for some higher notes. The sound becomes breathy and the vocal folds shift to a different plane." (P01, track 14)

Another participant, who criticised some of Estill's postulates, does not see the plane as a defining physiological factor:

> "Plane doesn't matter either, because plane is merely the inevitable effect of shortening vocal folds where the back goes down and the front stays where it is, or lengthening in which case they come up." (P12)

6.4.3 How many dimensions? Weight and stiffness

P12 suggests a second dimension for the modal vocal folds vibration: in his opinion thick or thin folds exist in the context of short or stretched folds:

> "You can do thick folds or thin folds and also short or stretched. It's two dimensions."

The volume of the folds remains constant, therefore one would expect a reciprocal relationship: the more stretched the folds are the thinner they would be and vice versa. But in practice by making them stiff or unstiff a singer can entrain weight: if the folds are relaxed (unstiff) vibration begins at the more flexible edges of the folds with its full amplitude, which decreases while it travels from the edges to the less flexible parts of the vocal folds body, this is also called light folds; if the folds are stiffened though, the whole body of the fold bounces in vibration, thus making them heavier:

> "You can have short folds or long folds, and you can have them thick or thin, within the context that they will always be thinner if they're stretched. But you can have relatively relaxed stretch folds which will give you a thin reedy voice. Or if you are looking at a counter tenor who's still producing that. You can make it a much more massive sound." (P12)

P12 employs the term *stiff folds* to denote adjusting the weight of the vibrating part of the folds. This is contrary to Estill's use of the term, she associates stiff folds with breathy sound and a closure that is not tight.

P12 uses *lax folds* as opposition to *stiff folds* to mark relaxed or unstiff state. P15 agrees with this terminology: she criticises the Estillian term stiff folds pointing out that stiffening a muscle means activating it.

> "For me one of the key aspects of falsetto is that the vocal folds are lax. I know in the Estill model they call it stiff folds, but physiologically that's not actually a good word to use, because stiffening muscles means activating them. So the vocalis is lax, and that's what I hear, but you can still tilt the thyroid with that to get a little bit more stretch on the vocal folds and possibly a little bit more contact between them." (P15)

At the same time there seems to be a contradiction between P12 associating lax, relaxed folds with head voice, good closure and vibrato, while P15 relates it to falsetto. Falsetto in Estill terms suggests no vibrato, a breathy tone and vocal folds barely touching. Also, while for P15 *lax folds* seems to be a better term for Estill's *stiff folds*, P12 uses *stiff* and *lax* as synonyms for Estill's *thick* and *thin* folds (but in opposition to *long* and *short* folds):

> "I don't think you get stiff vocal folds much with breathy sounds." (P12)

6.4.4 Relationship to other descriptors

In our chosen Track 24 vocal folds vibration mode has been rated consistently between mixed and thin, averaging at mixed thinner. Notably, the confidence here was lower than usual:

> "... with that much twang present, it sounds thick, but we know it isn't, when we measure it with instruments. So I would have said it was mixed thinner, because it is still a thickish sound, but probably being created by impedance rather than by the muscularity of the folds. And I would say the confidence is probably about 3. These are very misleading sounds." (P10).

It seems that the presence of twang and non-linear aerodynamic effects it produces have serious implications on the sound and on experts' ability to decode physiology. Another participant even suggested that falsetto might have been employed:

> "It could be falsetto with lots of twang." (P15)

She goes on, explaining how twang can be produced in falsetto:

> “The twang is made around the epilarynx and although you get that interaction with the folds so you’re going to get more resistance, to increase the resistance, you could still have relatively lax folds. I mean you can definitely twang in falsetto. “ (P15)

In a different song, Track 18, P10 again mentions twang as the confusing factor in terms of register and thickness of the folds. He goes as far as to suggest that even singer’s perceptions are affected by the back pressure created by narrow AES:

> “It’s a low head register, even though the folds are a little bit thicker. I think that’s because of the back pressure from the twang. I would say that she feels like she’s in chest.” (P10)

Another physiological descriptor that plays an important role in understanding registration is the larynx position. P01 remarks on Track 18:

> "It is not at all that high. Because of the high larynx, you get a feeling it is very high, but it’s not." (P01)

In a different track (Track 14) there is an opposite situation: the larynx was lower than expected for the pitch, impacting the intonation:

> “In the second snippet the singer changes from chest voice to falsetto, though only for some higher notes. The sound becomes breathy and the vocal folds shift to a different plane. At the same time, though the pitch gets higher, he does not take his larynx higher, he leaves it where it was, thus making the larynx position rather low for the pitch. This is probably the reason why it feels like his intonation is not quite exact. If his larynx position were neutral, I would hear the intonation as exact.” (P01)

P01 also mentions thyroid tilt affecting the rating of vocal folds thickness:

> "I stay with mixed, because there is such a strong tilt" (P01)

whereas P10 warns of exactly this kind of connection:

> “The problem you get is when you start thinking of those modes of vibration versus things like thyroid tilt. Your traditional Estillian singing teachers will be taught that the more you tilt the thyroid the thinner

> the folds get. And it's not true. Because when you take into consideration forces like impedance, you take into consideration how much of the muscles inside the vocal folds that you can actually engage at will as well, ain't necessarily pull the tilt away. So you can tilt as hard as you can and still sing very heavily. So, people often think volume, thickness of vocal folds versus thyroid tilt and it's just not terribly accurate." (P10)

On the example of Track 30 participant P12 points out a reciprocal connection between all vocal source descriptors: subglottal pressure, transglottal airflow, length and thickness of the vocal folds:

> "it's not as thick as it would be in a full chest vocal bit. But it's a compromise, you're injecting more thickness and shortness into the gesture. Because everything there is a reciprocal relationship to airflow resistance and pressure, and you can't change one without at least one of the others being changed. Yes, we learn how to play with that." (P12)

6.5 Thyroid tilt, cricoid tilt, cricothyroid visor

Thyroid cartilage tilt and *cricoid cartilage tilt* are typically Estillian terms taken from larynx physiology and used in isolation as building blocks for physiological settings of various vocal sounds (see Section 2.1.3.2). Singing teachers with the Estill system background operate with these all the time. Teachers who are not familiar with Estill terminology often haven't even heard of such things. In Estill approach there are well-defined acoustic correlates for both tilts. Medical professionals, though familiar with the terms, do not necessarily relate them directly to the same acoustic outcomes. Moreover, doubts have been expressed whether the cricoid cartilage can be tilted at all or whether the two cartilages can be controlled in isolation:

> "Cricoid movement and thyroid movement as isolated movements which affect sound production to me are secondary responses as a result of other things. So I would discount both the thyroid and the cricoid cartilage as physiologic responses in terms of the impact they have on the sound quality." (P11)

One participant with a medical/physiological background sees both cartilages as a single system (with some degrees of freedom) which he calls a cricothyroid visor:

> "But you need to remember that it isn't actually the position of the thyroid, it's the position of the thyroid relative to the cricoid. " (P12)

Thyroid tilt was rated by 7 and *cricoid tilt* by 9 out of 13 participants. Specifically for cricoid tilt we also used a perceptual alternative: yelling in the sound. Jo Estill employed cricoid tilt as a building block of belt with the above acoustic function. Yelling was a description everyone could relate to straight away.

Agreement for both thyroid and cricoid tilt was very poor.

6.5.1 Thyroid

In our chosen Track 24 *thyroid cartilage tilt* ratings span over the whole scale, showing no direction whatsoever. It seems that in the given vocal apparatus setting it is very hard to say whether the thyroid is tilted. Though participants agree that the vocal folds are thin, not all of them are convinced that this automatically implies a thyroid tilt:

> "Thyroid cartilage tilt. I mean – who knows? Who knows? I would not like to say. It could be vertical for all I know. The thing is, there is more than one way to thin the vocal folds. So, you know, I'm going to say I don't hear tilting. I'm going to say vertical." (P15)

While P15 is not very confident about this rating, another participant, P10 is convinced that the thyroid has to be slightly tilted, he infers it from other descriptors which he considers salient here, such as *AES*, the larynx position and the subglottic pressure. Elsewhere in his interview he gives an even more detailed account of the interplay between the thyroid cartilage tilt and the vocal folds mass (see his quote in 6.4.4).

The differences in participants' views on the relationship between register and thyroid tilt have been discussed above. While thyroid tilt influenced the vocal folds vibration mode ratings for some, there is also an inverse influence, when the tilt is concluded based on the register:

> "The thyroid tilt is vertical, otherwise he would abandon falsetto" (P01, Track 14)

This is a reflection of an Estillian view, where thyroid tilt is associated with thin folds, and falsetto is a different physiological mechanism, where vocal folds shift into a different plane. Another participant indirectly contradicts this view, for her

the thyroid can be tilted in falsetto in order to get more stretch on the vocal folds and a better contact between them (see P15's quote in 6.4.3).

6.5.2 Cricoid

Our experts are sceptical about the cricoid tilt mechanism as it is described by Jo Estill. There are doubts that tilting cricoid is physiologically feasible, and whether it has ever been observed empirically. The validity of the study on which Jo Estill based her conclusions about cricoid tilt is called into question:

> "Of course we know that physiologically there aren't any muscles to tilt the cricoid as such." (P15)

> "Cricoid tilt – there is no strong evidence to suggest that happens. And the study that looked at it was really poorly execution study. In examining hundreds larynxes during high intensity vocal manoeuvres we never once saw any consistent change of shape that would have suggested the cricoid is tilting. ... I don't know a single clinic, at least in this country, that could say yes, the cricoid is tilting, whereas you can see the thyroid, the view on camera changes." (P10)

So why was Jo Estill so keen to introduce such a controversial building block to her system? The reason for it was that she was looking for a physiological explanation of a particular aspect of belting: a yell in the sound. And our experts agree that this aspect is relevant and can be heard or even measured:

> "And we know when we hear it, when it becomes yell-like."

> "I do think we can measure this. We certainly can measure this yell-like thing acoustically.I do think we can measure it endoscopically by looking at the dimensions of the larynx. But I think it takes a little bit of training for the ear to really identify." (P10)

Assessing our Track 24 P10 remarks:

> "In this first example (snippet 24_1) there is no yelling. In example two (snippet 24_2), the same singer, there is a snippet at the beginning, where it does change. It's moving into this yell quality , but it's ... to do with vowel change. Vowel shapes absolutely impact on what's happening in the larynx." (P10)

P15 even suggested to change the question about cricoid tilt in our rating procedure to the question about presence or absence of yelling in the sound.

In no other context is the cricoid tilt building block relevant:

> "So for me the cricoid cartilage tilt would only be relevant if I heard someone I thought was belting. And for me that would be specific for a part of the vocal range." (P15)

Our experts are confident that there are some physiological changes which go along with the yell-like sound. They know it from their practice:

> "There is definitely something going on in relationship between the thyroid and the cricoid, in my opinion, when people are belting. And I actually teach people to achieve this changing the head position." (P15)

> "That is something that we're doing when we're belting giving us shorter but still relatively thick vocal folds. We've got this tacit shortening, some way or another, and that's how we're doing it." (P15)

Some even looked at it empirically:

> "What we did find, talk to Julian McGlashan or Catherine Sadolin on this, the dimension of the epilarynx changes in a slightly different way to twang. You see a squaring off just above the vocal folds, the false folds do start to get engaged. " (P10)

> "I think Lisa Popeil had a look at the ... belting ... flouroscopy and she saw thyroid tilting backwards. So anything that changes that relationship between thyroid and cricoid is going to adjust the length and tension of the vocal folds with that; but whether the cricoid cartilage actually tilts in the way that Jo Estill described it I'm skeptical. But I will say that something is going on." (P15)

In our chosen Track 24 the cricoid cartilage tilt descriptor displays no agreement between raters and ratings span the whole scale. Due to the fact that just three participants rated this descriptor for the track nothing can be said statistically about it. We notice though that the rater who hears the strong tilt is very unsure about his estimation, while the other raters suggesting a slight or no tilt are more confident. They point to the vowel shape as a defining factor for the cricoid tilt (see P10 quote in 6.5.2). P11 also points to vowel quality as a salient feature in this singing fragment.

6.5.3 Cricothyroid visor

One of our participants was very explicit about the physiological mechanism behind both vocal functions – thinning out the folds and yelling in the sound – which Jo Estill explained through thyroid and cricoid tilts. In his view it was not helpful to talk about the thyroid and the cricoid cartilages in isolation. His term of choice was the cricothyroid visor.

The cricothyroid visor is a rigid, cartilaginous structure in which the two cartilages – the thyroid and the cricoid – are double-jointed at the back, one on each side. It is attached flexibly to the trachea. The arytenoids sit on top of it. It is opened and closed by the cricothyroid muscles.

> "I use the word 'visor' because it's clearly a clamshell, or closure of the cricothyroid mechanism. " (P12)

The cricothyroid visor is the mechanism which is responsible for lengthening or shortening of the vocal ligaments through its opening or closure. When it is closed, it lengthens and therefore tensions and thins the vocal folds. This is what Jo Estill described as the thyroid tilt. Because the visor is a lever, this configuration has a mechanical advantage and can be sustained. When the cricothyroid visor is open, it pulls the arytenoids back and inwards and the arytenoids pull the back end of the vocal folds down, which shortens and thickens them. This is what Jo Estill called the cricoid tilt.

> "What is actually happening is the back of the vocal folds is being dragged backwards when you close the visor. ... what is important about that is it tensions and lengthens the vocal folds, but it doesn't look like it because they stay where they are, and the front of the thyroid drops away." (P12)

There are various physiological configurations which could lead to shorter vocal folds. The common denominator is that the folds are shortened when the visor is opened:

> "Shortening the vocal folds. There are two ways to do it. One is to open the visor. In fact, you can pull the thyroid back if you like, and that will also shorten them a bit. ... Actually the length of the folds is also dictated by the posterior cricoarytenoid muscles because if they collapse, then the whole thing tips." (P12)

> “You can also do by simply tightening up the posterior cricoarytenoids.“ (P12)

It’s much easier to open and close the visor with a slightly raised larynx by adjusting the altitude of the thyroid. With the low larynx it is rather the cricoid that would move to adjust the opening of the visor, though this configuration is rather rare. No matter which cartilage actually moves, the defining feature is the opening of the visor, the position of the cartilages in relation to each other.

> "- Does it make sense to talk about thyroid tilt as such at all?
>
> - Yeah because people can hang it on something that Jo made popular. But you need to remember that it isn’t actually the position of the thyroid, it’s the position of the thyroid relative to the cricoid. Whether it’s that, or that doesn’t matter. What happens is this clamshell closure which pulls backwards." (P12)

Shortening the vocal folds would normally make them thicker, but as discussed previously, there is another dimension of importance here: the mass or the stiffness of the folds. It depends on how much vocalis activity is involved. Though tightening/stiffening the folds the singer can change their mass. When stiff the whole vocalis body tends to move, while with lax folds only their edges will vibrate. This notion of vocal folds stiffness differs from Estill’s term of stiff vocal folds, which implies breathiness. Estill also talks about the "plane" of the vocal folds; for our participant the plane is defined by the opening of the visor as well - see his quote in 6.4.2.

6.5.4 Discussion

The main conclusion of the above discussion should be that though the vocal functions proposed in the Estill model are real – thinning out the vocal folds for thyroid tilt, yelling in the sound for cricoid tilt – the actual physiological mechanisms are much more varied, allowing for multiple strategies, for some degrees of freedom but not others, involving other parts of the apparatus apart from the thyroid and the cricoid:

> “When you read about the different ways the thyroid can be tilted: you can have the thyroid moving down to the cricoid, or you can have the cricoid moving upwards towards the thyroid, or you can have the thyroid

> moving forwards – that's three possible movements, and you can read about this in someone like Dickson and Mary Dickson, and Tom Harris also talks about this. So, if you think that there are three different ways of achieving of what we think as simply a thyroid tilt, it would not really be surprising if there was more than one way of achieving belting." (P15)

Also employing thyroid and cricoid in isolation seems improbable. It seems that this aspect of the Estill model, while addressing real vocal function, is far too simplistic in physiological terms. We need a better understanding of mechanisms for thinning and thickening vocal folds and a better terminology to describe it.

6.6 Larynx height

From the descriptors defining the basic physiological setting of vocal production we now move to those related to the resonance volume and configuration, affecting the sound after it had been produced. These descriptors include the position of the larynx, the AES, velum and tongue.

Larynx height relative to a neutral larynx position (see discussion of its definition in 2.1.3.2) was the only descriptor which was equally well understood by all the participants and displayed a tendency to agreement in its ratings. There were no misunderstandings or questioning of the term. It seemed to be self-explaining for the participants and it was apparently part of everyone's vocabulary. The agreement about the ratings confirms that participants in fact have a similar understanding of this physiological trait and its function. Quite often larynx height was also mentioned as one of the most salient factors.

This consensus about the larynx position was reflected in the ratings and discussions of Track 24. All experts rated the larynx to be high or very high. They also gave a high confidence for their ratings. (See Table 5.9).

P12 rated the larynx to be high, though not very high:

> "I think it's in a high laryngeal position but I think she could go strictly higher if she had to." (P12)

P15 relates the larynx height with its size: people with a small larynx may sound as if they take their larynx up high:

> "Position of her larynx is very high. And my confidence is also very high there. Or maybe she has an extremely small vocal tract as well, those two are of course related." (P15)

High larynx can produce a bright sound (Track 30):

> "R: What do you think he does to get this bright sound that he has?
>
> P12: Well, it's very high. I am sure it is a high larynx.
>
> R: You had said before that he has a small space...
>
> P12: I don't think he is a big man with a big larynx."

We even get a direct confirmation that larynx position is easy to determine:

> "I think the larynx position is very easy to identify, it's what's changing around it that is sometimes harder." (P10)

6.7 AES – the size of the vocal tract

Quantitative analysis shows that there have been a clear tendency to agreement among our participants about the ratings of *AES* (aryepiglottic sphincter) width. Yet in contrast to larynx height neither the definition of this descriptor nor its function or physiological mechanism have been agreed upon or easily understood. Joe Estill introduced aryepiglottic sphincter as a building block of twang, that (together with a higher larynx and lower velum) give the vocal sound a bright colour and a piercing quality. Aryepiglottic sphincter is part of the epilarynx, allowing for narrowing or widening the hypopharyngeal volume. It is part of the supraglottic vocal apparatus shaping the sound after it was produced by vocal source, though, as we discover in the discussion, non-linear effects affecting the source via AES adjustments might be taking place either (see Section 2.1.3.2 for more details on AES).

The term *aryepiglottic sphincter* was understood exclusively by those participants who were familiar with the Estill model, which clearly proves that it is "endemic" and is not associated with the stated functionality outside the Estill system:

> "I can think of no one who deals with the AES or tilt of either the thyroid or cricoid except Estill's followers. I find, therefore, that these are not universally accepted terms, understood across the profession." (P11)

At the same time, given the huge influence of Estill's work on today's teaching and research in the West, we thought it was justified to offer the term to the experts we interviewed.

One of our participants, well informed about Estill's work and terminology, insisted that the physiological mechanism of the described vocal function was not related to the sphincter but was a result of vocal source adjustments.

To explain the concept to those not familiar with Estill, terminology descriptions like *twang*, *narrowing*, *small space*, *bright sound* were used. For some participants it was clear from the interview, which term they preferred. Unfortunately not enough effort was made to define these terms, to differentiate between them and to track which of them led to what kind of understanding. At the same time, these terms were generally well received and understood, virtually no one had any questions or difficulties in rating it. Those fluent with the Estill system would sometimes use these terms as a synonym for the AES descriptor, in fact they even call AES the *twanger*:

> "I'd give confidence 3, because I don't know, when I'm in falsetto and I bring the twanger in, whether one can hear it."
>
> "I am not sure whether one can hear the AES setting in falsetto, given the breathiness. Changing the vocal tract form in this way would probably make no difference in falsetto." (P01)

As we have seen, there is an agreement about the values of the ratings which indicates that there could be some kind of shared knowledge behind this complex of terms.

For some participants this parameter, described by a variety of terms, was the most salient in Track 24: for P07 only twang was relevant, while P02 called it narrowing.

Let us now analyse in more detail the vocabulary our experts used when talking about the AES descriptor and the physiological mechanisms they believed were responsible for the agreed values of the descriptor.

6.7.1 Vocal function of AES – P12's concerns

Jo Estill ascribed the aryepiglottic sphincter the function of narrowing the hypopharynx, which would make the resonance space in the vocal tract smaller and the sound brighter, with a weaker fundamental and stronger harmonics (see Section 2.1.3.2).

Participant P12, a medical expert on vocal physiology with 45 years experience in the profession, was critical of this notion. He was convinced that brighter sound

was in the first place a function of voice source, not the resonance. He acknowledged that narrowing of AES exists, but attributed the brighter vocal sound to shorter and thicker vocal folds.

> "R: My understanding is that she [Jo Estill] tried to describe it, to say, well, we use our aryepiglottic sphincter to add a particular quality to the sound.
>
> P12: What makes it is the shorter, thicker vocal folds."
>
> "[Jo Estill] was confused about producing Broadway belt type. You're making things shorter and thicker there, and that is a reflection of the sound source. It's not a resonant thing." (P12)

He also mentioned middle constrictor in relation to brightening up the sound. In his opinion the role of the middle constrictor was major compared to the AES:

> "Well if you are looking to make formants and things, then what you use is your middle constrictor." (P12)
>
> "[Jo Estill] would actually see if she were not so busy that there is the middle constrictor at the same level as the aryepiglottic folds." (P12)

In the expert's view vocal source was clearly primary to middle constrictor in producing twang: the larynx is raised and the cricothyroid visor is tilted forwards. The role of the middle constrictor is to create a narrowing of the vocal tract which may cause non-linearity in the dynamics of the transglottal airflow:

> "It may be that you are actually reinforcing a phase difference sub- and supraglottic, but at the same time you have got to play that against the inevitable shorter, thicker vocal folds." (P12)

According to the participant, under some rare circumstances, mainly in sopranos on high notes, narrowing the aryepiglottic sphincter can have an effect on the brightness of the sound. He attributed it to the mentioned mechanism of creating a phase difference in the transglottal airflow which can appear in a narrowed vocal tract.

But this is not something he would see often. He said that middle constrictor would be more commonly involved, but on high notes the palatopharyngeus muscle activates which allows for the sphincter to be used. And the physiological mechanism used by tenors, who commonly employ a bright sound colour, would again be entirely different.

He called tenoring an "arrested swallow": when we swallow, the larynx goes up and the hyoid goes back; the epiglottis then reflectorily folds over and seals the larynx, protecting the trachea and the lungs from food and drink getting into them. In order to produce high notes tenors have to raise their larynx, but in order to prevent the epiglottis from closing they keep the hyoid forwards of the rim of the thyroid. High larynx, the hyoid bone in front of the thyroid and as a result a narrowed palatopharyngeal volume are the characteristics of the tenor gesture according to the expert.

> "If you keep the hyoid in front, the epiglottis can't go back. It's locked up because there is this thing called the hyoepiglottic ligament, which will actually keep the thing locked. If you're trying to close, there's a small muscle called the thyroepigloticus. That will narrow the front and you get the tenor gesture. It all looks very hyperfunctional when it is over, but it actually does a very interesting thing. It reinforces the anterior third of the vocal folds." (P12)

And the mechanism that keeps the hyoid bone in front of the thyroid cartilage, not allowing the epiglottis to close and the gesture to result in swallowing, includes the middle constrictor pulling back on the thyroid:

> "So you're holding the hyoid forward, the geniohyiod is the only muscle there is that can do it, and you're pulling the thyroid back with the middle constrictor. Out of that, you get some ways to produce a singer's formant. You get the anterior third pressing together really tight, and you get this very familiar tenor gesture." (P12)

The middle constrictor thus has a two-fold function here: to pull back the thyroid and to allow the singer's formant production. The ventricular folds are a bit pressed in, and so are the anterior third of the vocal folds. According to P12 it's the front of the vocal folds that makes most of the noise even though there's little of the vocalis muscle there.

Nowhere in this description of the tenor gesture does AES play any role.

6.7.2 Small space, narrowness

This terminology seems to come closest to the Cantometrics knowledge of narrowness. This is what the participants said about Track 24:

> "A narrow, a small vocal tract" (P15)
>
> "whether it is AES or a pharyngeal constriction, the space is very small" (P11)
>
> "Regarding open and closed sound, it's this narrowing or the feeling of narrowing" (P02)
>
> "I definitely hear a narrow, small vocal tract to produce that sound quality." (P15)

While most participants readily accept this term, P12 disagrees with it preferring to use *volume* instead. He associates small space with small bodies: birds, kids are small and very efficient, they can make a lot of noise with their sound source. A small person also has a small vocal tract. When we adjust our supraglottic structures to produce a brighter vowel, it is the lip radiation, the tongue position that are changed but the length of the vocal tract remains the same. His term of choice is the *volume of the supraglottic vocal tract.*

Summing up, *narrowness of the supraglottic vocal tract* seems to be a good approximation of what the participants mentioned in this context. It refers us directly to the Cantometrics terminology, where narrow sound is one of possible ratings for the *vocal width* parameter. At the same time narrowness still includes ambiguity, whether it is lips, or tongue, or jaw, or AES, which can lead to confusion and to quite different acoustic results.

6.7.3 Brightness

What is a dark/bright vocal sound? When the term was suggested to the participants we meant by dark a stronger fundamental and by bright stronger other harmonics. This description is based on partials values, which are determined by the supraglottic space, thus being a resonance characteristic. Two of the participants directly addressed the question of this terminology. Participant P12 said:

> "Dark is when you back everything. Short thicker folds, you pull everything back and you drop everything very low. It makes a very powerful noise, but it needs to say that because you've created this back sound by dilating everything up it tends to be nasal, you can't keep it tight." (P12)

So the larynx is lowered, the pharynx widened, the tongue lowered. The resonating space is enlarged. The lowered larynx makes the vocal folds shorter and thicker.

This configuration is loosely related to our definition above, since a larger resonator better amplifies lower partials.

This is how participant P11 defined dark/bright sounds:

> "bright vowels meaning smily wide vowels and dark vowels meaning kind of narrowing inside" (P11)

Interestingly, for her, bright vowels are associated with wide and dark vowels with a narrowing. "Smily" refers to the lip form, while "narrowing inside" can only mean the palatal configuration. If the lips are protruded the mouth has a form that is closer to a tube, amplifying lower harmonics. Spread lips change the form of the mouth to become a narrow slit. This effect can also be illustrated by the vowel formants: in "smily" vowels like eee and iii the second formant is high, far above the first, while in round ooo and uuu it is very close to the first thus reinforcing the fundamental.

We see that although these three descriptions of dark/bright sounds seem quite different, there is still a common direction, with lower harmonics amplified in darker vowels and higher harmonics in brighter vowels.

How do singers produce bright or dark sounds then? The AES is only one of the factors. One participant associates brightness with the tongue:

> "I know that brightness is controlled by the tongue. When the tongue is high, there are higher overtones resonating, so that one does not need a lot of twang and still has the higher overtones." (P02)

Another participant mentions high larynx:

> "- What do you think he does to get this bright sound that he has?
>
> - Well, it's very high. I am sure it is a high larynx." (P10)

Our chosen Track 24 is a good example of a bright sound. P10 implies high larynx and either AES or the tongue, or all of them:

> "Well, this is the tricky thing, even though the sound is thin and it's not that the formants are particularly high in the perception, something has got to be put on that bright cut on the top, so it either would be larynx and tongue, or the AES, or all of the above. I think the larynx position is very easy to identify, it's what's changing around it that is sometimes harder. I still think that AES and the tongue are kind of

question marks here. One of them has got to be helping out, but I'm just not sure which one." (P10)

But then P12 brings in vocal source:

> "...there are the timbral effects that you can get at the sound source that are also important and I think it's probably slightly thinner fold – still modal but it is a thinner fold than might otherwise be used, with not absent vibrato, but light and tight." (P12)

In particular he notes a good closure of the vocal folds with very little breathiness:

> "It is actually associated with very little breathiness, and really quite good closure even though it's a small percentage of the vibratory cycle, and certainly there is a whole load of bits that you get with the resonance as well." (P12)

He explains the mechanism of achieving such a good closure by narrowing the hypopharyngeal space which causes cross-phase in transglottal airflow and therefore a snap of the vocal folds:

> "Brightening up the sound certainly has to do with the snap of folds back. If you are closing that [AES], I would allow that the phenomenon of the sound source that may well be producing a phase difference in the transglottic airflow, in which case you actually get a push backwards. It's like doing it with bottles or singing through straws and that sort of thing. If you can change the air pressure above the vocal folds and below it so there is a phase difference, it rises above as it falls below, then you get a snap back and it may be that closing up the egress is helpful for that." (P12)

So far thickness of the vocal folds, good vocal folds closure, larynx height, AES, tongue and lips have been mentioned, though participants' focus varied. It gives us quite a diffuse picture of the physiological mechanism of vocal brightness. P12 confirms the confusion:

> "I am honestly not sure I would want to venture in what the changes are specifically with bright or less bright relating just to the resonance. I would give myself at least a [confidence of] 2 or less on that." (P12)

P11 insisted that it would be easier for listeners to agree about the dark/bright characteristics of vocal sounds than about physiological configurations. She conducted a study in which she investigated this agreement as well as the physiology behind the sounds. Seven women sung dark and bright vowels in three different vocal qualities (chest, mix and head). The agreement among evaluators about which vowels were dark and which were bright was very good. At the same time, the physiological strategies used by the singers to make the vowels bright were all different. The common trend was to make the space in the vocal tract smaller, but they all did it in different ways:

> "Some people were narrow, some people brought the larynx up, some people actually brought their velum in, they did all different manoeuvres inside to shrink the space. Some of them were very visible in the mouth, we did a mouth evaluation. In some of them the mouth remained the same, but the sound changed because the inside structures changed. The aryepiglottic sphincter changed in most people, but not in everybody." (P11)

In P11's experiment AES did play a prominent role though it was only one aspect of a large number of various physiological adjustments the singers made. This experiment tied together the narrowness of the supraglottic vocal tract we mentioned in the previous subsection and the brightness of the sound. P11's findings about a good agreement among human listeners on the dark/bright characteristics was in line with our quantitative results, displaying a good agreement between participants on the AES ratings. At the same time P11 showed that a wide variety of physiological strategies were used by the singers to produce bright sounds, which is in accordance with a rather diffused picture of physiological mechanisms of brightness that emerged from our interviews.

6.7.4 Constriction, contraction, edge

P11 used the term constriction specifically in the context of AES discussion on Track 30:

> "...you could say that there's constriction in this sound. But there are many places for constriction to happen, and the aryepiglottic sphincter is not the only one. So confining the constriction to just that particular area, which is a very Estillian idea, limits the kind of description you can

> see, because this is constriction in the base of the tongue and the back of the mouth, which hasn't got anything to do with the aryepiglottic sphincter." (P11)

In Estill approach, constriction is mentioned in conjunction with the retraction or contraction of false (ventricular) vocal folds; aryepiglottic sphincter which also belongs to the epilarynx is considered separately, following the main Estill idea of isolating physiological building blocks.

P11 continues, introducing the term contraction, which is deliberate and controlled as opposed to a habitual constriction:

> "It would be fitter to say that this is a constricted sound but I wouldn't call it constricted, I would call it just contracted. In other words I think that this is a sound that's probably typical for this person and therefore that's the only sound that he knows, which is opposed to tightening something on purpose. In my definition of physiologic response there are two different kinds of constriction: one is habitual and unconscious, which cannot be affected in a direct way, and the other one is kind of contraction done for musical purposes which is deliberate and controlled." (P11)

Later, when prompted directly to choose between the terms AES and contraction, she introduced *small space*:

> "You would have to say that this is a very tight place, so the space inside is very small, however you get there, it's a small space. If it's aryepiglottic sphincter or it's just pharyngeal constriction – it's small. But she's comfortable in this small space." (P11)

While Jo Estill does not relate constriction to AES, but only to the false vocal folds, other participants mention it in conjunction with the middle constrictor or a pharyngeal constriction in general. Yet another participant talks about constriction in the context of "a reduced freedom of the vibrating portion of the vocal folds" (P10). A discussion with P01 took place during the interview where summarising the behaviour of the false vocal folds and the middle constrictor under the term constriction was considered. We eventually set up for the term *pharynx form*, rejecting constriction as too general. AES was not mentioned in any of these discussions. It seemed that the term *constriction* was too ambiguous to be related directly to AES.

The Complete Vocal Technique (CVT) by Catherine Sadolin (Sadolin 2000) employs their own, somewhat idiosyncratic terminology. At the same time, it is a very popular approach in Europe and more and more vocalists and teachers are familiar with the terms. One participant, a CVT adept, characterised Track 24 as *edge*:

> "In terms of CVT I would call this edge. It's the fourth mode that needs a lot of twang. "

While twang can be found in other CVT modes, *edge* is louder, has metal in the sound and more overtones.

6.7.5 Discussion and relationship to Cantometrics

AES was introduced by Jo Estill as one of physiological building blocks to explain/construct twang. It is closely related to twang and is even called *twanger* by some. While a number of participants referred to AES in this context and in the context of their own research, one participant insisted that the physiological mechanism behind twang is based on the vocal source in the first place, not on the resonance. He acknowledged though that some narrowing could take place and even produce a non-linearity in the transglottal flow dynamics, causing a snap back (a fast and tight closure) of the vocal folds, and therefore a bright sound colour. Other participants attributed the bright colour to *AES* as well as to a high larynx, the position of the tongue, a pharyngeal constriction (e.g. middle constrictor), the lip form, etc. Some referred to *AES* as a form of constriction, while others didn't.

As we have seen, the *AES* descriptor in terms of its definition is more than any other descriptor based on a set of different terms and notions, in other words, a lot of confusion. At the same time, mysteriously, participants consistently agreed about its ratings. This could be a lead to an uncharted territory in music psychology and in voice perception. In fact, one participant, a highly experienced singing teacher, teaching and performing in a large number of music styles, directly confirms this: in her view, it would be much easier to hear the opposition between dark and bright vocal sounds, and to teach others to recognise it, than to actually identify which physiological strategies were used to achieve these sounds. Here we have made a full circle and returned to where we started: to a perceptual descriptor, like the Cantometrics *vocal width*, that is taught by verbal descriptions and musical examples. Though the term is different, the principle is the same. As mentioned elsewhere, it is quite complicated to design an experiment to confirm or refute this

phenomenon in general human listeners. Yet if we receive further leads pointing in this direction it might well be worth pursuing in future research.

Our participants' ratings do confirm the hypothesis in P11's experiment, displaying an overall agreement about the values of AES/size of the vocal tract, though we have seen that physiological mechanisms described by the participants were varied, diffuse, sometimes even contradictory. We arrived at the term *narrowness of the supraglottic vocal tract* as a good compromise between participants about the mechanism of achieving bright sound. It directly refers us to the Cantometrics description of "narrow sound", one of possible values of the *vocal width* parameter (compare Table 2.4). There is no indication though that this narrowness is related to physiological tension (hyperfunction) or the impression of tension in the sound, as was assumed by Alan Lomax. This is supported by the fact that most participants did not use the term constriction in relation with AES or brightness (compare Section 2.3).

Constriction could be related to the Cantometrics description of "tense, squeezed" sound; if we use it, it would provide us with another convenient link to the Cantometrics *vocal width* parameter, similar to the one between the narrowness of the supraglottic space and the "narrow sound". As we have just shown though, the notion of constriction is too ambiguous: there are many different places in the vocal apparatus where constriction can happen, and its effects on the resulting sound can be very different; constriction can be habitual or intentional, again producing a range of sounds and impressions. Physiologically speaking we use the term constriction in relation to particular muscles or groups of muscles (middle constrictor, aryepiglottic sphincter), but then there is always the balance of constricting and lengthening muscles in every movement. We have seen that none of these specific constrictions spans the whole picture of producing a bright sound. Adding to the confusion, the term constriction is used differently and ambiguously by non-medical professionals, e.g. singing teachers. All this makes the term as well as the general notion of constriction an unsuitable candidate for a formal ontology of vocal production.

6.8 Tongue and velum

> "Velum – I don't think the velum tells us whether anything is tense or relaxed. It's to do with resonating qualities. And, actually, I would say similarly with the tongue." (P15)

For many of our experts both tongue and velum do not affect the basic physiological setting of the vocal apparatus. Like for P15, they influence the resonating qualities. They can be used to change the volume of the vocal tract and to increase brightness of sound:

> "I know that brightness is controlled by the tongue... When the tongue is high, it gives you more high overtones that co-vibrate, so that you don't need so much twang and you still have lots of high overtones." (P01)

Both of these characteristics are in turn affected by the phonetics of the language the singer speaks and sings (see quotes in 6.8.1).

Not all see velum and tongue as equally unimportant for the given setting. For P02 velum plays a very different role compared with the tongue:

> "The tongue is for me an additional component, it can affect the sound negatively, but, in contrast to, say, velum, it would not change the character of the sound." (P02)

And for P14 the significance of the tongue depends on the voice:

> "The tongue is important, but not the most important. To me. It is important in some voices but then not very important in some others." (P14)

As there is no clarity on the significance of both descriptors, there is also little to no agreement about their ratings throughout our experiment.

6.8.1 Velum

Velum is the flexible bit of the palate which separates the nasal and the palatal cavities. It can be raised at will to close up the nasal cavities from the palatal side, or it can be relaxed so that there is a channel connecting the cavities. It can be lowered even further widening the connecting channel. This gives the vocal sound the distinct nasal quality.

> "it's another way of making the vocal tract smaller narrowing the sound." (P15)

We use the term velum in the Estillian sense to refer to the closure or opening of the nasal channel adding nasality to the sound. This obvious vocal function is often

overseen by singing teachers who refer to soft palate exclusively in relation to its other function – raising and tensioning the anterior part of the soft palate to assist with resonance and help manage the passagio (see Section 4.4). P15 pointed out this difference:

> "when you put mid, are you saying that ... it's a nasalised sound or are you simply using a more loose kind of singing teacher type language which is 'I've raised my soft palate or I've lowered it'?" (P15)

All participants were comfortable with our understanding of velum and its function (though in one case there seemed to be a misunderstanding of the scale, as can be seen in Figure 5.5.1). Yet participants' opinions on the significance of this descriptor varied greatly. P15 and P10 considered it "the icicle on the cake" – not relevant for the basic vocal setting:

> "Don't care about the velum. You know, the velum is just the icicle on the cake. It doesn't really tell us much about the phonation mode. It doesn't tell us what the singer is doing, it is just carrying the sound. That's very much my take. You can have a high or a low velum with any other setting that you've gotten here. It is relevant and it is useful to know but it's not the thing that really tells me what the singer is doing in terms of vocal production." (P15)

Others did see its significance for the physiological setting: P07 and P01 name velum as a salient factor on a number of tracks. P02 stresses its importance in comparison with the tongue:

> "The tongue is for me an additional component, it can affect the sound negatively, but, in contrast to, say, velum, it would not change the character of the sound." (P02)

It's not just the significance of the parameter velum that our participants' views diverge on. They mention two factors which in their opinion might lead to differences in rating the parameter. The first concern is raised in relation to our chosen Track 24. It is the difficulty of rating physiology for vocal sounds one has no experience of making. The effect of empathic listening can get in the way and lead the rater astray. One of our experts who has experience of singing in Southeast Asian vocal traditions is not very confident in her rating of Track 24. She notes:

> "Velum mid, confidence 3. When it goes towards India or Southeast Asia, I'm not quite sure. When I imitate these sounds I have to take the velum down. I know singers who claim it is not a nasal sound. But I don't know how good they feel it." (P01)

Another factor related to the familiarity with the culture is the phonetic content of the language in which the singer sings (and speaks). Two participants mention the difficulties of determining the velum setting without the knowledge of the language:

> "Velum is tricky with the stuff that's in a different language" (P10 in relation to track 22)

> "It's difficult to say with the velum because of the linguistic patterns." (P15)

All these differences make it plausible that in our experiment we have found no agreement between our raters on the values of velum opening.

6.8.2 Tongue

For many of our experts the position of the tongue is not defining for the general physiological setting of the vocal apparatus:

> "Almost always the tongue is the least relevant descriptor." (P02)

Its position is often arbitrary and hard to determine without actually looking inside with the camera:

> "I would say that the tongue adjustment is arbitrary, simply because the tongue really operates in isolation" (P11)

> "Tongue position is always inferred. Because you can't really hear it. When we look at it acoustically, it doesn't make a gigantic amount of difference. So a mid tongue position or a high tongue position, if the singer is highly skilled in twang, isn't going to make a gigantic difference in terms of what you perceive." (P10)

While in Western classical singing teaching practice it is common to advocate for a low tongue, P12 stresses that it is a myth, that a low tongue is beneficial for singing:

> "You only have to look at where the three tenors hold their tongues variously... none of them has got a low tongue." (P12)

In his opinion it is the volume of the vocal tract that is crucial, while the position of the tongue is changed with each vowel.

> "Oh for God's sake, it's a volume thing, and it depends on where the thickness is. But it's not stable, and it's nothing right, you have to modify it in any case for every vowel you make." (P12)

In this respect our Track 24 presents a notable exception. Interestingly, there is a very good agreement on the tongue position, though this descriptor displays little agreement in general. It seems that the overall configuration of the apparatus in this vocalisation only allows for a certain tongue position. The participants are also unusually confident about their tongue ratings. Some of them actively listened to the tongue:

> "The tongue is ... probably high" (P02, this participant explicitly said in his interview that he can generally hear the position of the tongue vey well).

Others inferred the tongue position from the general configuration of the vocal apparatus in this recording:

> "The tongue is high, can't be any other way." (P01)
>
> "I think the tongue would have to be high." (P10).

And another participant nails it:

> "Because of the rest of the musculature around there the tongue is lifted in the back. I don't know if you have a category for that, but the tongue is up, not down. It's the side effect of everything else." (P11)

While most participants consider the tongue position not salient for the perception of the singing on this recording (for P02 all descriptors were salient apart from the tongue), P15 actually points to the tongue position as a salient factor, probably because it is indicative of this rather extreme configuration of the vocal apparatus and is also easy to determine here.

In relation to the tongue participants specifically pointed out its potential to inhibit the effectiveness of vocalisation:

> "I am very sensitive to the tongue, when the tongue or something else on this axis interferes with the sound." (P01)
>
> "The tongue ... can affect the sound negatively, but ... it would not change the character of the sound." (P02)

For example in Track 32 everything in the physiology points toward an open sound for P01, but it does not sound open. She blames the compressed tongue for it, that is pushed in the space above the vocal folds and hinders the free flow of sound.

Here is another example: in the second snippet of Track 18 P01 assumes that on the vowel "a" the tongue is lowered. It then triggers other things, such as the activation of the middle constrictor. It can be heard anywhere in the song, where the singer changes to “a”:

> “In my opinion this is the conflict between the high larynx and the tongue that is a little bit too flat for that larynx position, so it is a bit unbalanced. ... The most salient for my perception is the middle constrictor, this phenomenon: a high larynx, a tongue that is a little bit flat, and this triggers the constrictor.” (P01)

This example points to the importance of mutual interaction between different parts of the vocal apparatus, when tiny changes in one of them affect the others and thus influence the whole physiological setting. When discussing Track 24 P10 suggested several physiological components that could be responsible for the brightness of the sound: larynx position, AES or the tongue (see his quote in 6.7.3). P11 confirms the complexity of interaction, bringing in other components:

> “behaviours of the tongue are almost always coupled with impact from something else, like jaw, or the swallowing muscles, or lips and mouth muscles, unless the tongue is severely limited, in which case the languaging of the singing would sound distorted.” (P11)

P15 gives an example of a negative impact of the tongue position on other components and thus on the whole setting:

> “I’m hearing a big jaw space and a low flat tongue. And a low larynx, he could even be lowering his larynx via the tongue. That’s the reason why to me it sounds unstable.” (P15)

6.9 Other descriptors mentioned by participants

In an open coding procedure we collected vocal physiology and vocal production characteristics mentioned by our participants for which no systematic ratings were collected.

6.9.1 Soft palate

Soft palate is a wide spread term used by many singing teachers. In Western teaching tradition singing teachers are often concerned with raising the soft palate to improve the sound and the technique. They sometimes ask their students to imagine they are yawning or are biting in a sour apple: these prompts are used to help students activate and raise their soft palate. The Latin term velum and the English term soft palate are in fact equivalent in physiology.

The complication around velum stems from two facts. First, the function of raising the soft palate is not well defined or understood. Johan Sundberg showed that raising soft palate helps to achieve the singer's formant, which is one of the ultimate goals of Western classical singing education. Jo Estill employs raised soft palate (called *head and neck anchoring* in her terminology) to produce a belt sound. Yet there is no general explanation as to what exactly it does and why it is so important. Second, velum has another function in vocalisation: to separate the nasal cavity from the mouth. The velum's softest and most flexible posterior part (ovula) can be raised to close the channel between the nose and the mouth; or it can be lowered to open that channel, this gives the sound the nasal quality (see Section 2.1.3.2).

The velum or soft palate is a flexible, non-homogeneous tissue, which is stiffer at the edge closer to the hard palate and softens towards the ovula. In its anterior part it can change its form only slightly – when tensioned it is raised somewhat, attracting metaphors like vault or sail. Its softer posterior part can hang down, can be brought up to close the nose channel, or can be set in vibration by a stream of air, like in French "r".

Jo Estill used the term velum to refer to the latter function – opening and closing the nasal channel. Therefore for her low velum means nasal sound and high velum means absence of nasality. This is what one of our participants was referring to when pointing out the two different functions of the velum (see P15's quote in 6.8.1).

There is a parameter in the Estill model that is related to raising the sail. It is called *head and neck anchoring*; it is a combination of posture and an activation of a whole group of muscles, which include muscles triggering the tension of the soft palate. It is used to add power to speech and singing (Estill et al. 2005a, p. 109). *Head and neck anchoring* is used as a building block in both belting and operatic sound.

Discussing velum with one of our participants whose first language was not English

we were looking for terms to differentiate the setting and function of raising and lowering the soft palate from the nasality function. We used words like palatal curtain, sail, dome, vault. The participant preferred to call it tension of the curtain:

> "Not the dome, but the tension of the curtain. It can be completely relaxed, like a hanging cloth. It can be slightly tensioned, or a bit more. A strong tension is possible, but this is already wrong singing technique." (P08)

She explains the function of soft palate in speech, pointing out when it is tensioned reflectorily:

> "In speech the curtain is relaxed and lowered, it is only raised reflectorily by very strong emotions, such as fright, yell, wail. ... Like when you cry out "Aaauuu!" in the forest. When you need a high range in a speech situation, than it is raised automatically." (P08)

While most singing teachers are concerned with helping singers to activate the soft palate, P08 quite remarkably talked about deactivating it. She primarily teaches a non-Western singing style which is based on speech mode and where a raised soft palate is not always beneficial. According to P08 some singers have a very sensitive, flexible palatal curtain and have difficulties keeping it relaxed, therefore they cannot retain the speech mode in singing. That happens often in singers who get to the top easily, such as classical tenors.

> "Who has the sail naturally always working, for them it is hard to switch it off. The sail does not know what to do in a speech mode." (P08)

The Western academic singers, she stresses, raise the sail and loose the speech mode. But those who want to stay in speech mode, if their soft palate is too active, they have a barrier to change from head voice to chest, to relax the sail. But for the change from chest to head it needs to be activated. It plays an important role in mixed voice:

> "The curtain needs some tension for the register break, this is very important. For example, you can't do much in a female mixed zone without an active soft palate". (P08)

To summarise, P08 stresses the importance of tensioning the anterior part of the soft palate to manage the passagio and to create a mix of chest and head; and of its

relaxation to obtain or retain speech mode in singing. She missed this descriptor in our ontology and suggested to add it.

This participant was not the only one who used the term. P01 even named it as one of the most salient for Track 18:

> "For tension [the most salient descriptors are] soft palate, velum, subglottal pressure." (P01)

6.9.2 Middle constrictor

There are three pairs of muscles in the pharynx whose primary function is to contract the pharynx in order to transport the food into the oesophagus (Figure 2.1.15). One of them in particular, called the middle constrictor, is situated exactly at the level of the larynx. Three of our participants discussed it extensively in the interviews; one of them explicitly suggested to add it to the ontology. The views of the three participants on it seemed to differ significantly. Two participants stressed that not much has been researched or published about the middle constrictor and its role in singing yet.

> "That's a really new one, we don't have enough data to actually say with any confidence that that's consistent." (P10)

P10 came to talk about the middle constrictor in the context of discussing the bright tone in Track 20. He hypothesised about what could have produced that bright edge, mentioning AES, high larynx and the tongue. These were all components present in our ontology. And then he mentioned another possible source of brightness – the middle constrictor. In his view it has an acoustic response similar but slightly less than the AES.

> "Where this would be really complicated, is – because the AES has a particular acoustic response, there is also another structure in there which is a middle constrictor and crico-pharyngeus, which Diana Harris of the Harrisses would call the ring of confidence. And that was something that we see a lot in folk and opera singers but also in pop singers. And it gives you an acoustic boost that's just slightly under what AES tends to give you. So it's a less harsh version." (P10)

He hears it in Track 16 as well, stressing that it is not the AES giving it the brightness, but the constrictor.

In his own research he observed the middle constrictor in action with clinical instruments. He explained that it's at its most active when a person cries:

> "But where you really notice this particular combination of middle constrictor and crico-pharyngeus is this narrowing of the whole, well it looks like the narrowing of the whole pharyngeal wall, is when you really cry, cry very hard, when you're using very very criey sound. That comes in automatically, and also brightens the sound slightly. So it's not always the AES that's giving you that brightness. " (P10)

He goes into more detail about the state of research on middle constrictor:

> "But I would guess that it's only the Harrisses clinic and my old clinic who would even acknowledge that that's the thing. Because they have ... a very long bit of research they did; we found it separately in a long bit of research we did; we started talking to each other, we did a joint piece of research and we were all very pleased that we felt we found something. We've never got round to publishing it. " (P10)

Another participant who also did his own research on middle constrictor, sees its function in producing twang and belt. He disagrees with Jo Estill on the physiological mechanism of belt. In his view belt is the result of vocal source adjustments and not as Estill claims of activating AES. In his view AES only plays a marginal role. What could be of more significance is the middle constrictor. It could be facilitating non-linear aerodynamic effects which create the resonance, the twang:

> "... a purely resonant thing happens at about the same level which is the middle constrictor, and also palato-pharyngeus narrowing." (P12)

Generally, he remarks that middle constrictor is used to create formants (e.g. the singer's formant):

> "Well, if you are looking to make formants and things, then what you use is your middle constrictor." (P12)

For example, tenors use it to produce the specific tenor configuration of the vocal apparatus:

> "What you have to do is tilt and untilt the thyroid relative to the hyoid. It's quite interesting because as soon as the hyoid goes back in a raised

> larynx, the epiglottis folds right over and seals the larynx. Now, tenors will raise to the point because they're still hanging on with the middle constrictor to make this constriction to give themselves a singer's formant. But the tenors will keep the hyoid; it's not out there in print much at the moment." (P12)

Middle constrictor is also paramount to producing belt:

> "You get mostly middle constrictor, but depending on the gesture, and if you are going to the top you get a bit more of the palato-pharyngeus at the back coming out, so you get an arrangement around the side that can use that sphincter [the AES], and that's where it happens." (P12)

P01, an Estillian, gives her own view on the middle constrictor. For her, when the constrictor is activated, it is often a sign of excessive effort. Its activation is connected to AES:

> "In my case, when I add twang, my constrictors are also activated. But it does not add brightness, it rather sounds somehow lumped." (P01)
>
> "My middle constrictor switches on when I use AES. I am now at the point where I try to isolate it. There is an interaction [between the middle constrictor and the AES], but I know how it sounds when I relax the constrictor. It becomes more open." (P01)

For this participant letting go of the middle constrictor is associated with openness, large space in the throat, or the Cantometrics *vocal width*:

> "That has always been my understanding, with conventional singing teachers as well as now [with the Estill system] that when someone is supposed to have lots of space in the throat, it is about the constrictors. That they are not active, the muscles, but loose and you get more space through that." (P01)

She analyses how the middle constrictor is activated by the singer in the second snippet of Track 18 as opposed to the first snippet and it makes the configuration unbalanced. This is for her the most salient feature of the snippet:

> "... the main difference is that on the vowel 'a' the tongue is lowered, and it triggers other things, such as the constrictor. It can be heard anywhere in the song, where the singer changes to 'a'. ... In my opinion this is

> the conflict between the high larynx and the tongue that is a little bit too flat for that larynx position, so it is a bit unbalanced." (P01)

P10 in turn offers some critic of the way how Estillians understand middle constrictor:

> "They very often talk about the middle constrictor when they are hearing curbing, but that's not it, it's the base of tongue, which ends up eliciting that particular sound quality. So, I think there is more research here to be done and somebody's got to publish something before we know for sure." (P10)

6.9.3 Head position

One participant insisted that head tilt was important and should be added to the ontology. While discussing Track 24 she stressed that the singer was probably singing with her head lifted:

> "My guess is that if we saw the second lady here, the Taiwanese woman, she could very well be singing with her head in a lifted position – that was the impression that I got." (P11)

She goes on to explain that a slight head tilt is desirable:

> "The head position, that is to say the head is tilted slightly up, makes a great deal of difference in everything else, because it helps the inside structures adjust." (P11)

According to her, a slight tilt is common for most non-classical singing styles in the West. Head tilt helps the larynx to rise adjusting to higher pitches.

> "Almost all sounds that are not classical – the head slightly lifts up. Like this. And that has to do with the release of the constriction inside. If someone is trained to keep his head always straight and never move, when you ascend in pitch, you reach a place where the larynx needs to rise and move, and it can't." (P11)

Another participant also noted that she uses a slightly lifted head position to teach belting. In her analysis she also referred to head position:

> “I think there is thyroid tilt. He might use cricoid tilt on the top notes. I wouldn’t like to say, if I was looking at the head position and actually able to see the movement of the larynx I could be more sure.” (P15)

This last citation points to the main reason why including head position in our rating procedure could be problematic: when you see the singer, you can use their head position as a clue for physiological configuration; but it’s very hard to infer head position from only hearing the singing.

6.9.4 Pitch

A number of participants, though not all, checked the exact pitch or pitch range of singing in the snippets, and that informed them about other characteristics:

> "The knowledge that it is around A, in a higher range, proves to me that she cannot be totally relaxed." (P10)

> “If they were singing a B flat below middle C that would be pretty low for a falsetto register.” (P15)

Some participants suggested adding pitch to the ontology because it was informative of the physiology involved. In one case a participant completely changed her ratings after checking the pitch. In Track 18 a singer sounded very high for P08 and she based her ratings on that knowledge first. But then, after checking the pitch and finding it was an octave lower than she assumed, she had to revise her judgements. P01 who was more experienced in Southeast Asian traditions and had sung in a similar style herself, could identify the situation more easily (see quote in 6.4.4).

Pitch seems to be a useful addition to our ontology, since not only is it informative, it is also very easy to determine and unambiguous.

6.9.5 Vibrato

5 out of 13 participants explicitly mentioned vibrato in their analysis. While listening to Track 28 P08 was informed by the vibrato that there was a good vocal fold closure and, though the singer dropped the pressure, it was not a breathy sound. P05 noted that vibrato was absent in the list of terms, and, while she didn’t hear much vibrato in the examples, it could be an important aspect. Yet when she was asked whether presence or absence of vibrato would change her perception of the singing as being wide/relaxed or narrow/tense, she considered it for a moment and concluded that her perception wouldn’t change.

P14 also asked for vibrato to be added. He commented specifically on Track 16 that vibrato was important, and he heard vibrato on other tracks as well. P15 and P12 characterised the singing on Track 16 as wobbly, which had a negative connotation. P12 suggested that it was related to the age of the singer and the accompanying vocal problems:

> "it's very, very ageing, low wobble, because that's what happens to us all in old age. And if you can keep your vibrato tight, you will sound much younger than you deserve to." (P12)

Here P12 gives a more detailed account of how vibrato is produced and what is a good vibrato. In his opinion a good vibrato does not involve changing pressure or larynx position. It is all resonance adjustments, which may give an impression that pressure and intensity fluctuate. He also stressed that there are many different strategies for producing vibrato:

> "There is as many ways of producing vibratos as singers, almost. Most of the vibrato that we find acceptable is nothing to do with the 'hahahaha' sort of pressure around. You can change pressure so you get air pulses, you can change pitch so you get a wobble, 'waaaaa', which is dreadful. That's to do with age as well so we don't associate that with youth." (P12)

> "Actually the most successful vibratos that I see are mostly all resonance features, not pitch, not intensity. They're neither. All of them have been described as vibrato, and most of them are hideous. The one that I think is most successful is where people are messing slightly with the formants and the resonance so it seems to be pitched, and it seems to be a bit intensity as well. But mostly it's neither, it's your message. If you look at a good vibrato, mostly you'll see the most of the activity is in the pharyngeal musculature on the base of the tongue, rather than anything violent going on within the larynx." (P12)

6.9.6 Volume

P07 requested to have volume in the ontology. According to him, in Complete Vocal Technique (CVT) they do not work with vocal source as a separate entity. As Johan Sundberg shown volume is directly related to subglottal pressure (Sundberg 1987). Therefore, in some sense, CVT does work with subglottal pressure, but

through the lens of volume. This substitution fits into the general picture: CVT, though informed by Estill's physiological approach, consciously works with acoustic descriptions of sound and not with physiological building blocks. Catherine Sadolin (2000) developed a sophisticated categorisation of vocal sounds, including vocal effects, using her own vocabulary. CVT has become very popular, particularly in the world of contemporary commercial music, it seems that many people have been able to recognise their perception of vocal sounds in Sadolin's terminology. And in her world listening to the volume makes more sense than hypothesising about subglottal pressure.

6.9.7 Articulation 1: jaw and mouth corners

The main physiological components of articulation are tongue, jaw and lips. They determine the formants and the shape of the spoken and sung vowels as well as the construction of most consonants, thus defining the content of our speech. We included the tongue in our ratings to investigate whether the participants would display similar opinions about its average position in our examples and whether any commonly understood vocabulary related to tongue would emerge. We decided to exclude the jaw and the lips due to the fact that the raters had no access to the singer and could not see his face, and we felt that rating them would be too hard. Also, since articulation changes with every syllable, we thought that changes in jaw and lips happen on a time scale too small for our experiment, where we needed several seconds of physiologically stable singing (see Section 4.4).

As we have seen, not a trace of clarity emerged from rating the position of the tongue, and we felt confirmed in our decision to exclude other articulation components. Yet some of our experts did miss the jaw and the lips in the ratings. P11 referred to Johan Sundberg's work and to the fact that the jaw affects the formants:

> "The jaw position changes the formants. If everything else is the same but the jaw opening and closing is more, ... there is a difference in that sound and the only thing that's different is that I'm moving my jaw down more on each of the syllables I'm singing. That's the only difference. So I think that matters, because it changes the length of the vocal tract and that changes the formants. So, the jaw position – that's Johan's teaching, not mine." (P11)

P15, while listening to Track 16, heard a big jaw space and noticed that it was missing. She related jaw space to the perception of big or small space in the vocal

tract:

> "There is a possible here that you have not included, which occurs to me. I hear him singing with the jaw space. So I'm hearing a big jaw space and a low flat tongue. ... Since you've got your overall impression, which is narrow or wide, you could argue that the impression of big space or wide space or small space includes jaw." (P15)

P05 mentioned both jaw opening and the lips – mouth corners either spread or rounded. She was convinced that she would be able to rate both of them in our examples. What she seemed to imply was that there might be habitual or stylistic factors defining the average position of jaw and moth corners:

> "For example the country and Western guy, I would say that he has spread articulation, it's completely speech-like, he wouldn't mind at all to sing "iiiiii", "eeeeeee", "aaaaaa", whereas in the first Chinese example there was very very little of that kind of articulation. It was more rounded." (P05)

For P02 though the jaw was secondary, determined by other physiological factors. The same was true for the tongue, the jaw being the more significant factor of the two. Stiff jaw and tongue for him usually indicated that there is tension in the system already:

> "The tongue, like the jaw, has a secondary function. These are for me two additional parameters, in this order of importance: first the jaw, then the tongue. A stiff jaw is the worst, it makes the tongue stiff as well, and this is for me a secondary [compensatory] tension. This tension appears when there is tension already, then the jaw and the tongue would in the worst case add to that tension. And even if you take away that tension in the jaw and the tongue, the original tension can remain. That's what is often done in a wrong way: they reduce the tension in the jaw and the tongue, but the source of the tension is still there. As long as breathing, subglottal pressure and transglottal airflow do not change, there is no need to relax the jaw and the tongue. That's my phoniatric opinion." (P02)

For him it was subglottal pressure, transglottal airflow and larynx position that defined the tense or relaxed state of the vocal apparatus. Relaxing these would lead to adjustments in the jaw and the tongue:

> "I claim that as soon as subglottal pressure and transglottal airflow and, say the larynx position are optimised, the position of the tongue and the jaw will be positively affected. You cannot develop singing technique just with the tongue. You can adjust it afterwards, to improve the sound and the intelligibility of the vowels further, but it won't make your voice any different." (P02)

6.9.8 Articulation 2: phonetics and vowel shape

Articulation is not limited to tongue, jaw and lips, it includes all physiological components related to resonance. In the previous subsection we mentioned the habitual or stylistic conditions defining the average jaw opening or lip rounding. Yet what affects the configuration of the resonance body the most is the phonetics of the language that the singer speaks and sings. P08 explicitly pointed out to me that this should be included in the ontology:

> "You should talk about phonetics and how phonetics – the shaping of sounds – affects the timbre, the vocal production and all the physiological parameters we talked about." (P08)

P01 engaged a lot in pointing out differences between vowels in her analysis:

> "I think it fluctuates a bit depending on the vowel." (P01)

> "I hear it on the "aa", but not on the "ää"." (P01)

She mentioned that she was very sensitive to the tongue, which probably explains her attention to different vowels, since the tongue position is largely, though not completely, defined by the vowel.

For example, she observed less tight vocal folds closure on "a" vowels in track 16 and explained on which auditory perceptions she based her conclusion:

> "On the A's the tongue goes down and the airflow increases, it's not completely closed then." (P01)

While track 16 comes from Southern Europe, track 18 is a gamelan piece from Southeast Asia. Though the tracks and the singing in them are completely different, P01 again stresses the difference between "a" vowels and other vowels (see her quote in 6.9.2).

P08, when analysing Track 34, heard in the lower notes of the first verse a particularly beautiful, richly resonating sound, but this effect disappeared in the second

verse. The reason was the change of the vowel from a round O-like sound to a narrower one:

> "What a huge difference! He began the second verse, the phonetics changed and the timbre on the lower notes is really different. The phonetics was narrower in the second verse and it was all sung in the same timbre. And in the first verse the timbre, the colour was very different at the bottom notes. It's just the vowels, their phonetics! And look, the timbral colour changes. And in terms of physiology it is all the same, no change." (P08)

She confirmed that all other physiological factors remained the same. Therefore for her the fragment was physiologically stable. It was just the change of the vowel that produced quite a big acoustic difference. Here she touched on the reason why we did not include articulation changes in general in our ratings: the main disadvantage was the different time scale (see Section 4.4.2). Vowels and articulation usually change within milliseconds, while for our experiment we were looking for physiologically stable fragments that were several seconds long.

When she said that in terms of physiology there was no change it was not really precise – there would have been changes in lips and tongue to change the formants (which define what vowels we hear). What P08 might have meant was that apart from these adjustments in the mouth there were no changes at the source or otherwise in the vocal apparatus configuration. A possible explanation to the acoustic difference she heard could be the effect of vowel tuning, where a singer implicitly tunes one of the vowel's formants to coincide with a harmonic, thus enhancing this harmonic and the overall vocal output. Possibly this effect occurred for one of the vowels but not for the other.

While for P08 with the vowel change all other parameters remained unaffected, other participants did find that changing the vowel led to adjustments of other physiological parameters. For example, on Track 16 P01 heard that the vowel A triggered consistent changes in the setting. Everywhere in the song where the singer changed to A she heard less brilliance/more noise and she also assumed that the middle constrictor was triggered (see her quotes in 6.9.2 and 6.9.8).

Similarly, in track 24, P10 noticed that in the second snippet a yell quality appeared which was absent in the first snippet. He attributed this difference to a change in vowel shapes. He went as far as to say that vowel shape would affect not only the resonance but the larynx itself (see his quote in 6.5.2).

For P11 the vocal quality of the vowels (which includes their shape) was the cornerstone of her analysis of singing, because they contain the most information on the vocal production:

> "The driving force of any analysis to me is what I hear in the vowels. The vowels' sound carries the most energy, so I'm always listening to any kind of sustained quality on the vowels, because most information I need is there. So that would always be the primary driver. Especially when you can't see the person, when all you have is your ears, you have to pay a lot of attention to the vowel sounds." (P11)

She mentioned that in her teaching practice, in order to diagnose technical weaknesses in particular passages of her students she would make them sing on sustained vowels, then increase the intensity/volume slightly. If the student is capable of doing it, it usually means that the technical setup of the vocal apparatus is stable and works well; if not, it highlights the weak points very effectively. In our experiment the experts had no access to the singers and for P11 it was particularly important to analyse the vocal quality in the vowels she heard.

What is this auditory information P11 extracts when listening to the sustained vowels? She was quite specific about which characteristics of vocal production she listens to: from nasality and breathiness, over dark vs bright to pitch and volume (see her quote in 6.10). For example, she explains how she infers the physiology related to small or large shape of the vocal tract (see our discussion of the size of the vocal tract in Section 6.7.2) for the vowel shape:

> "The shape in the vocal tract could be construed to be "small" (a more intensified sound with higher amplification enhancement) or "large" (where the larynx rests low in the throat and the fundamental is amplified) by evaluating the type of vowel sounds being sung, as the vowel carries most of the qualities we perceive in sustained sounds." (P11)

And then P11 goes on to explain that decoding physiology is based on hearing the above characteristics in the sustained vowels:

> "So if I hear a certain kind of vowel, and a certain kind of tone in the vowel, and a certain kind of quality in the vowel sound tone, then that tells me all the things we just talked about: where is the velum, where is the larynx, how much airflow, whether the person is comfortable. So the first answer is always going to be the sound. How do we know what the

physiology is? I've listened to the ingredients of the aural spectrum." (P11)

Overall, the vowel shape, though it could not be reflected properly in our current rating format, seemed so important, that we included it in Chapter 7 as a common theme. That chapter summarises our findings about the topics that were raised by many participants and which we thought crucially influenced their thinking and hearing of the recordings we offered, and therefore their ratings. In other words, these common themes are starting points to the explanation of the disagreement between the raters in our study.

6.10 Conclusions

Analysing experts' comments on their ratings and their understanding of ontology descriptors fully demonstrates the complexity of the subject and the variety of reasons and concepts underpinning their decisions. Some common themes have emerged while we analysed this complexity, and they constitute the subject of Chapter 7.

As we have seen, participants did not agree about any vocal source parameter. Therefore, the general physiological settings in our musical examples remain a mystery. Even most basic facts, such as whether pressure and airflow are directly related; which measures of vocal folds are important and the mechanisms of adjusting them; the existence and the nature of registration – were not agreed upon. As expected, thyroid and cricoid tilts produced quite a lot of scepticism and were only rated by a fraction of participants. On the resonance side, position of the larynx seems to be easy to determine, which was explicitly confirmed by one of the experts. That was the only parameter in our ontology which showed a tendency to agreement and didn't raise even more questions. At the same time, velum or nasality of the sound, which would seem to be a clear and easy to determine parameter similar to the larynx position, didn't display any agreement. Position of the tongue was not well defined in the ontology, since there was no obvious, commonly used concept or terminology for it. We included it rather to explore participants' attitude to it, and, as expected, no agreement was found.

The biggest surprise that emerged from our analysis though was that there was in fact an agreement about one of the most mysterious descriptors in our ontology – the *AES*. The term that is not known outside of the Estill community, whose vocal function is not clearly defined, that produced the most discussion and controversy in

the interviews – it was one of only two parameters in our ontology about whose ratings participants were in a consensus. It was discussed in relation to twang, belting, brightness of the sound, vocal tract volume, constriction. One of the participants, a renown expert on vocal physiology, was vehement that the aryepiglottic sphincter, introduced by Jo Estill to explain twang, was not involved in the physiological mechanism of the acoustic results we were at. At the same time, participants rated this descriptor more consistently than any other and their confidence in their ratings was mostly high.

What that tells us is that there must be an acoustic phenomenon related to the above complex of concepts that is easily identifiable by experts through listening; and that this acoustic phenomenon is better agreed upon than the possible physiological mechanisms behind it. It is an indication in this particular case that physiology would not constitute the best choice of ontological approach for vocal production and that it might appear more fruitful to search for the kind of acoustic anchors which can be widely agreed upon, at least within one musical culture. Our participant stated:

> "We can perceive, auditorially, vocal quality (clear, nasal, breathy, pressed, noisy), vowel sound quality (toward forward/bright or toward dark/back) and register quality (heaviness/lightness), as well as pitch, volume (decibels or SPL) and we can imply from the decibel level the subglottic pressure levels. The shape in the vocal tract could be construed to be "small" (a more intensified sound with higher amplification enhancement) or "large" (where the larynx rests low in the throat and the fundamental is amplified) by evaluating the type of vowel sounds being sung, as the vowel carries most of the qualities we perceive in sustained sounds." (P11)

The results of this study confirm that the experts who took part in fact agreed about a descriptor related to what she called sound quality (dark/bright sound). Unfortunately, we found no agreement about the other items on her list: register quality, phonation, nasality. What the participant suggested as the basis for an ontology is loosely reminiscent of Alan Lomax's Cantometrics approach: nasality, vocal width/vocal tension, volume, interval range, etc are among his parameters describing singing. This PhD originated from an attempt to deconstruct and reformulate the Cantometrics parameter of *vocal width*. A good agreement on *AES* and *larynx height* which both displayed consensus and both correlate with *vocal width*

(see Section 5.8) demonstrates that there must be some truth behind Lomax's approach in this particular case. That we did not see agreement on other descriptors such as velum (nasality) just underscores the complexity of the subject.

A number of terms were suggested as additions to our rating procedure, some of which can be determined objectively and included into our ontology, though none of them were suggested by a majority of participants. This demonstrates that our list of terms comprised the most relevant descriptors. Suggestions ranged from unambiguous, easy to determine characteristics such as vocal pitch to newly discovered physiological influences such as middle constrictor, where no consensus is expected among those who are familiar with it. Some terms are widely used by singing professionals such as soft palate or vibrato; others were related to articulation and were not rated due to the visual factor limitations (see Section 4.4.1). One suggestion – the vowel shape – was so important and so often mentioned though that we have channeled it into one of the common themes discussed in Chapter 7.

We present here an updated version of vocal production ontology that we compiled originally in Table 2.2. The updated version in Table 6.2 includes pitch and vibrato, that were missed by participants most often. It also seemed appropriate to separate Estill's head/neck anchoring into head position and soft palate tension. We substituted the Estill model specific term AES by the most widely agreed term *narrowness of the supraglottic vocal tract.* The flow phonation was dropped because it emerged that it was too specific to the Sundberg's discourse. Thyroid and cricoid cartilage tilts were rarely understood and rated outside the Estill informed community, their terminology needs to be questioned. Yet they reflect important functionality and acoustic outcomes; we therefore leave them in the ontology in the hope to find better terms in the future.

Table 6.2: Ontology of vocal production updated

descriptors	**physiological dimensions**	**range**	**scale**	**metrics**
subglottal pressure	subglottal pressure	low to high	5-point	interval
transglottal airflow	transglottal airflow	low to high	5-point	interval
phonation	phonation breathy	present/absent	2-point	nominal
	phonation pressed	present/absent	2-point	nominal

Table 6.2 Ontology of vocal production updated continued...

descriptors	**physiological dimensions**	**range**	**scale**	**metric**
	phonation neutral	present/absent	2-point	nominal
register	vocal fry	present/absent	2-point	nominal
	chest	present/absent	2-point	nominal
	head	present/absent	2-point	nominal
	falsetto	present/absent	2-point	nominal
	flute	present/absent	2-point	nominal
vocal folds vibration mode	vocal folds vibration mode thick to thin	thick/ mixed thicker/ mixed/ mixed thinner/ thin	9-point, NA	interval
onset	aspirate	absent/ occasional/ often	3-point	interval
	smooth	absent/ occasional/ often	3-point	interval
	glottal	absent/ occasional/ often	3-point	interval
thyroid cartilage tilt	thyroid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
cricoid cartilage tilt	cricoid cartilage tilt	vertical/ slight tilt/ tilted	5-point	interval
false vocal folds	false vocal folds	retracted/ mid/ constricted	5-point	interval
narrowness of the supraglottic vocal tract	aryepiglottic sphincter	wide to narrow	5-point	interval

Table 6.2 Ontology of vocal production updated continued...

descriptors	**physiological dimensions**	**range**	**scale**	**metric**
larynx height	larynx height	low to high	9-point	interval
soft palate tension	Soft palate tension	no tension/ slight/ strong	5-point	interval
velum	velum	low to high	5-point	interval
tongue	tongue height	low to high	5-point	interval
	tongue compression	present/absent	2-point	nominal
jaw	jaw position	back/ mid/ forward	5-point	interval
	opening	minimal/ mid/ drop	5-point	interval
lips	protrusion	no/slight/strong	5-point	interval
	spread	no/slight/strong	5-point	interval
head position	head position	straight/tilted	2-point	nominal
torso	torso anchoring	relaxed/ anchored	2-point	nominal
pitch	pitch range	C1-C7		interval
	pitch distribution	wide/ mid/ narrow	3-point	interval
vibrato	vibrato size	deep/shallow	3-point	interval
	vibrato frequency	wide/frequent	3-point	interval

The next chapter summarises the lines of discourse that emerged from the interviews, that were mentioned by several participants and that we felt played an important role in understanding the views of the participants and reflect the complexity of the task and the processes taking place during singing and during rating the recordings. These common themes give us clues about what could have led to irregularities in ratings or to disagreement between raters. They have consequences for various disciplines and are a source of new ideas and further research questions which we discuss in Chapter 8.

7 Meta-analysis, reasons for disagreement

Chapter 6 was dedicated to the qualitative analysis of the interviews performed with 13 vocal physiology experts for our study. Examination of experts' comments on their ratings and their understanding of ontology descriptors fully demonstrated the complexity of the subject and the variety of reasons and concepts underpinning their decisions. Some common themes have emerged while we investigated this complexity, and they constitute the subject of this chapter.

7.1 Differing interpretations of terminology

The high acceptance by the participants of the terminology used in our ontology (see Section 5.1) and at the same time a low degree of consensus about the ratings suggest that the different views were mainly based on terminological differences: while participants were familiar and comfortable with the terms, they quantified the parameters differently, without a tendency to agreement for all but two descriptors. Given the large discrepancies in the ratings it would be reasonable to assume that there are also differences in meanings projected on the terms. This is confirmed by our qualitative analysis of the interviews and the current subsection summarises our findings.

In some cases terminological differences are revealed by high confidence for diverging ratings, like for phonation. In other cases the differences are conceptual, like whether pressure and airflow are or are not inversely proportional. Sometimes terminology is misleading, like velum vs. soft palate. In a number of situations doubts were expressed about physiological mechanisms that were the basis for the terminology, such as cricoid tilt or AES. Also, in some instances such as tongue position quantification rules were not well defined.

Phonation

Despite obvious disagreements about phonation modes of singing in our samples, our experts mostly expressed high confidence in their ratings. It could therefore be assumed that the differences in ratings are due to different interpretations of the term phonation. For instance, we saw *pressed phonation* as a label of vocal health issues (see quote in 6.3.2). Often pressed singing was associated with inefficient, unprofessional or "incorrect" singing, some experts referred to this meaning (see quote in 6.3.2). And for others phonation was a non-judgemental characteristic of vocal production based on vocal source physics, related to other aspects of physiology (see P01's quotes in 6.4.2 and 6.3.2).

All these interpretations took place without any reliable measurements, based solely on experts' auditory experience. The differing interpretations of the term phonation aimed to describe different aspects of voice; we don't know at this stage whether this divergence also included genuine differences in the views on physiological reality behind the term.

The situation was complicated by the fact, that our participants were not familiar with the musical cultures from which singing examples originated (see quote in 6.3.2).

Sometimes, when confronted with vocal production from unfamiliar cultures, that work with levels of pressure unknown in Western music, some experts refused to call it pressed or doubted the usefulness of this terminology (see P05's quote in 6.3.2).

The unfamiliarity with the culture and the language adds another unknown to the freedom of interpretation. We discuss it in detail in Section 7.6.

Registration vs vocal folds thickness

Registration is probably the most contested, controversial concept in singing voice, and at the same time in practice it is used more often than any other aspect of vocal physiology (Section 2.1.3.2). All participants confirmed that they were comfortable with using the terminology of registration and that they employed it in some form in their daily work. At the same time the number of registers ranged from three to five; some called them head and chest, others said modal and falsetto, yet others used M0-M3 letters for registers; registration was either related to range or not related to it; some registers could mix according to some participants, others could not (see Section 6.4).

Jo Estill, probably the most influential singing teacher of the last century, from

whom we partly borrowed the terminology for our ontology, did not recognise registers as separate entities. She reformulated the notion of registration in physiological terms as a measure of thickness or thinness of the vocal folds, which could be adjusted gradually for any pitch (Section 6.4). None of our participants seemed to share these maximalist views, though P10 confirmed that for him registration is not related to range.

To cater for both notions of registration, classical and Estill-inspired, we included two descriptors into the ontology: *vocal folds vibration mode*, following Estill's definition, and the classical register with the *position within the register range* attached to it.

Our experts interpreted the relationship between the vocal folds vibration mode and the register descriptors in different ways. For P15 they were equivalent and the ratings could be mapped unambiguously. P02's interpretation was less straight forward based on his experience as a clinician, imagining the vocal folds vibration descriptor as one-dimensional and the register descriptor as a multi-dimensional stroboscopic picture (see their quotes in 6.4.1).

P14, whose own terminology included chest, head and falsetto, mapped Estill's thick folds onto modal or chest and thin folds onto falsetto (and not head voice). In his view head voice was a special configuration, a prerogative of Western classical singers (see his quote in 6.4.2). Therefore he considered it not appropriate to use the head voice category in the context of other cultures, for singers who were not trained in opera. Thick/thin folds opposition did not describe the particular configuration of head voice for him. Yet interestingly he did rate Track 22, a Balkan singer very different form classical Western opera, as head voice, while P15, who is much closer to Estill in her views, heard falsetto in this voice.

Velum

Velum or soft palate is a flexible, non-homogeneous tissue, which is stiffer at the edge closer to the hard palate and softens towards the ovula. Due to its heterogeneity it serves two different vocal functions. In its anterior part it can change its form only slightly – when tensioned it is raised somewhat, attracting metaphors like dome or sail. Its vocal function is to add resonance to the sound (it is part of the singers' formant's setting). Our experts mentioned it being indispensable in managing the passagio which is one of the main concerns of Western classical singing teachers, but is also important in any style that employs the mixed zone.

Its softer posterior part can hang down, can be brought up to close the nose channel, or can be set in vibration by a stream of air, like in French /R/. Its vocal function is adding nasality to the sound when the nose channel is open or removing it when it is closed. Jo Estill refers to this function when she uses the term *velum* and we followed her approach in our ontology.

Somewhat surprisingly we didn't see any complications in the interviews that were due to the double function of the velum. All our participants were physiologically informed and had no difficulty correctly referring to the nasality function. Therefore in this case we can say that the expected complication due to the terminological ambiguities didn't materialise. At the same time we had a misunderstanding in relation to the scale: we suggested to rate the position of the velum from low (open channel, nasal sound) to high (closed channel, non-nasal sound). Yet P14 seemed to have reversed the scale, probably rating nasality from low to high. Even when accounting for this misunderstanding we cannot explain the unexpected disagreement about nasality among our experts. This is the case where obvious terminological differences in regards to velum are not the cause of it.

Yet recent research addressed the polysemy of the term nasality and its relationship to the velum opening; this is where a deeper reason for disagreement on this descriptor can possibly be sought. Garnier et al. (2007b) provides a detailed analysis of the term semantics (including its close synonym twangy which is often more negatively connotated) as it is used by French singing teachers, displaying a cluster of distinctly different meanings:

> "... the polysemy of the "nasal" descriptor, sometimes related to forward sound placement, sometimes to lateral opening of the mouth or to nose constriction, sometimes to high-harmonics spectral reinforcement, and sometimes to a coupling between oral and nasal cavities by means of the velum." (Garnier et al. 2007b)

It could seem that these different aspects are functionally related, but this is not unambiguously backed by evidence. Fant, cited by Laver (2009), reported that lowering the velum leads to a significant drop of the first formant due to the resonating frequency of the nasal cavity and the effect of anti-resonances. Sundberg's (2007) results for one Western classical singer confirm this observation: F1 is attenuated so that the relative level of the singer's formant increased. This is in contradiction to Titze et al. (2001) and Steinhauer et al. (1992) who observed an increase of F1 in "twang". Garnier et.al (2007a) measured F1 reduction for a tenor

singer, an increase in F1 and F2 for a bass-baritone and no changes for another bass-baritone singer for nasal sounds. In fact Birch et al (2002) found no correlation between nasal quality of the vowel sounds as rated by a panel of experts and the presence of velopharyngeal opening in singers determined empirically by three different methods.

Our participants mentioned two related factors which in their view could have lead to uncertainty or disagreement in their ratings: being unfamiliar with the language of the song as well as with the vocal culture itself and vocal techniques employed. See Section 6.8.1 for quotes and examples; we discuss familiarity with the culture in Section 7.6.

AES

Jo Estill introduced *AES* as the physiological mechanism of twang: activating (narrowing) the sphincter helped to narrow the epilarynx and gave the vocal sound a piercing quality, shifting the harmonics spectrum to the right (see Section 2.1.3.2 for more details). The term *aryepiglottic sphincter* in connection with vocal function was understood exclusively by participants familiar with the Estill model (see P11's quote in 6.7). Given the importance of this functional aspect and the influence of the Estill model on singing education we decided to include the term in our ontology though, in the hope to discover alternative terminology during the interviews.

To explain the concept to those not familiar with Estill terminology we employed descriptions like *twang*, *narrowing*, *small space*, *bright sound*. We borrowed these terms from the interviewees who were familiar with the Estill model and the term AES. Yet this complicated the situation, bringing in all the perceptual and physiological variants of these phenomena. As a result AES turned out to be the biggest terminological mess in our experiment. It is therefore particularly surprising that AES was one of the two descriptors for which we saw a tendency to agreement. Below we summarise what we learnt from the interviews about our participants' views on the listed terms.

Twang was a very common substitute for AES often used synonymously by the Estill-aware participants (see e.g. P01's quote in 6.7). This is understandable given that Twang was the main vocal function of AES in the Estill model. Yet some participants denied a connection between AES narrowing and the acoustic outcome we call twang.

For most participants *small space* referred to a narrowed vocal tract, while P12 as-

sociated it with the size of the person. We finally converged on the term *narrowness of the supraglottic vocal tract* as a consensus.

P12 interpreted *bright* and *dark* in ways not related to *AES*: for dark sounds, alongside shorter, thicker vocal folds, you "pull everything back and you drop everything very low" (quote in 6.7.3); bright sounds were described via the elements like vocal folds thickness and closure, breathiness, vibrato (quote in 6.7.3). He mentioned resonance structures, of which AES would be part, but stressed that it is very hard to know which of them were involved and in what way (quote in 6.7.3).

Other terms have been suggested by participants such as *constriction*. A lengthy discussion took place with P01, an Estillian, on AES and middle constrictor being involved in producing twang. Looking for a term to describe them collectively, we considered constriction. In the end we decided against it – there were too many different interpretations of it in relation to vocal production. E.g. Jo Estill associates contraction of the false vocal folds with constriction. While P11 uses the words contraction vs constriction to differentiate between habitual and conscious tightening (quote in 6.7.4).

Please see Section 6.7 for more detailed discussion of these terms and for interview quotes.

7.2 Differing views on physiological mechanisms

While in previous cases there were differences in contexts and interpretations that caused disagreement between participants, in this section we look specifically at situations where participants' views of physiological or physical reality diverge. They include the dimensionality of vocal folds vibration and the belting mechanism.

Vocal folds vibration mode

Estill's definition of the vocal folds vibration mode between thick and thin is one-dimensional. P12 insists on two dimensions: length and weight (thickness, stiffness, see Section 6.4.3).

For Jo Estill the length and the thickness of the folds are inverse proportional: the longer the folds, the thinner they are. In P12's view it is the weight of the vibrating part of the folds that affects the vocal sound and is not directly related to their length. While the folds are thinned out when lengthened, and become thicker

when shortened, the weight of the vibrating part can be adjusted to counter these natural changes. This adjustment is achieved via the folds' stiffness: if the folds are relaxed (unstiff) vibration begins at the edges of the folds and decreases while it travels from the edges to the less flexible parts of the vocal folds body; if the folds are stiffened though, the whole body of the fold bounces in vibration, thus making them heavier. In this case it is not just the terminology that is different: in P12's explanation the physiological mechanism is conveyed in more detail. We can assume that Jo Estill was either unaware of this mechanism or didn't agree with this explanation of vibration mode: she used the term "stiff folds" to describe the state of the folds when they are in a partial contact only and therefore produce a breathy sound. Both P12 and P15 noted that this is not accurate, because stiffening a muscle means activating it; when the vocalis muscle is stiffened, the vibrating part of the folds becomes heavier.

AES

Jo Estill introduced *AES* – a supraglottic structure – to explain an important component of belting – the twang. One of our experts does not agree with her views on the role of AES. In his continuing argument with Estill P12 insists that physiological mechanism responsible for the acoustic result described by Estill has nothing to do with *AES*: in his opinion it is based on vocal source adjustments in the first place (see his quote in 6.7.1). Additionally, the activation of middle constrictor leads to non-linear aerodynamic process in which vocal source is affected by the resonator shape (quotes in 6.7.1, 6.9.2). See Section 6.4.3 for more details.

Other participants do not dispute the role of *AES* as such. Yet some are sceptical about the simplistic view Estill presents. For instance, P11 mentions that constriction in the vocal tract can be produced in multiple ways, many of which do not involve *AES* (quote in 6.7.4).

P15 indirectly supports P12's view on the non-linear aerodynamic effects playing a role in producing twang. She also mentions P12's second dimension of vocal folds vibration – vocal folds being lax means that a lesser, lighter part of the folds vibrates (see her quote in 6.4.4). This quote combines in the easy-flowing sentence a lot of different factors: non-linear aerodynamics, vocal folds resistance, vocal folds weight and falsetto. Such a mixture is quite complicated to confidently disentangle just by listening. Another participant confirms that you can't really hear whether there is twang in falsetto (see P01's quote in 6.7).

This leads us to the subject of the next section – difficult physiological configurations.

7.3 Difficult physiological configurations

In some cases it is just difficult: you can't really know what is going on in the vocal tract without looking at it. The exact physiology may be ambiguous or very hard to pin down just by listening. This is reflected in the low confidence of the experts in their ratings (see P01's quote in 6.7).

Non-linear aerodynamics

We mentioned the possibility of aerodynamic non-linearity in the previous section – the situation where the configuration of the resonance space affects the vocal source. While G. Fant's linear Source-filter model of vocal production has been useful in many applications, the non-linearity in vocal aerodynamics has attracted attention of researchers recently (Butte et al. 2009, Titze 2008). A good example of the presence of non-linear effects in our study was Track 24 from Thailand. P10 described the singing on that track as "very misleading sounds": it sounds thick, though not because of the actual thickness/muscularity of the folds, the effect is caused by a large amount of twang and the impedance affecting the vocal source (see P10's quote in 6.4.4). The situation is further complicated by the fact that the singer as well as an empathic listener – for instance one of our experts – would not notice the difference: whether the timbre is produced by thicker folds or if non-linear effects are in play (see quote in 6.4.4).

Larynx height

Track 18, a gamelan song from Indonesia, represented another striking case of mislabeling physiology in our study. One participant was tricked by the very high larynx to believe the pitch to be an octave higher than it actually was (see Section 6.9.4).

Tongue

While tongue position is not very well defined in our system, it is also a difficult physiological component to describe. It has many degrees of freedom, at the tip, the blade, the dorsum and the root. Participants' views on its role vary greatly (see

Section 6.8.2). P10 explicitly pointed to the difficulties in determining the tongue position:

> " . . . hard to know without actually looking inside with a camera" (P10)

While he was mostly confident about other descriptors, for the tongue he rarely gave more than 3 out of 5.

Special cases - Tibetan monks

To round up this section we would like to mention vocalisations which are physiologically speaking special cases, where non-conventional physiological mechanisms are used to produce sound.

One particular difficult case from our participant's practice came up in the interview: the Tibetan monks. These small-sized people sing with extremely low-pitched voices, producing very impressive, unique sounds. These sounds are too low to be produced by human larynxes in the usual way. So something else must be happening in their throats. One of P12's colleagues investigated. He was only able to determine how they make these sounds by means of high speed cameras that were inserted in the singers' vocal tracts as well as electroglottographs – there was no way to find out just by listening, therefore it would have been impossible for other participants to give informed ratings of such singing. (Luckily, we did not have any Tibetan examples among our music samples.)

What P12 and his colleagues saw was at the same time astonishing and simple. The frequency of the sounds was half the frequency of the vocal folds vibration. Moreover, they saw false vocal folds oscillating so that they sometimes reinforced the true folds' peak and sometimes canceled it out. This was the effect that created a very powerful sound an octave lower than what the true vocal folds produced (see Bailly, Henrich and Pelorson 2010).

Other examples of special cases in vocalisation include: yodel, overtone singing (Bailly, Henrich and Pelorson 2010), growl (Sakakibara et al. 2004a) or the aryepiglottic trill (Moisik, Esling and Crevier-Buchman 2010), Sardinian (and possibly Russian) very low bass singing (Henrich et al. 2006), Japanese Noh tradition (Yoshinaga and Kong 2012). There are vocal treasure chests collected by ethnomusicologists filled with examples of most extraordinary human vocalisations (e.g. Hugo Zemp's "Voices of the World"). Within the Western musical discourse Catherine Sadolin's Complete Vocal Technique addresses the question of vocal effects (Sadolin 2000).

7.4 Different physio strategies

In the previous section we investigated some difficult physiological configurations where our experts had doubts about their ratings, reflected in low confidence scores, or pointed out ambiguities and difficulties in rating. In this section we touch upon a related problem – when singers use different vocal strategies to achieve the same acoustic result. In contrast to the previous section where experts were aware of the difficulties, in this case the ambiguity of the physiological configuration may be concealed, in particular due to the mechanism of *empathic listening* we discuss in Section 7.6.

AES, bright sound, small space

We have discussed AES, sound colour and supraglottic volume in detail in Section 6.7 Here the focus is on different strategies singers use to achieve bright sound or a feeling of small space in the vocal tract.

One of our experts conducted a study in which she investigated how singers produced bright vowels as opposed to dark ones. P11 and her colleagues asked seven female singers to produce dark and bright vowels in three vocal qualities: chest, mix and head. The brightness of the vowels was evaluated by expert listeners: when the singers made brighter sounds, they were perceived by evaluators as being brighter or having a higher intensity energy in them, the agreement was good. All the singers made their vocal tract smaller to produce brighter sounds, but each of them did it in their own way, for instance moved their larynx higher, lowered their velum, changed the mouth shape, etc (see P11's quote in 6.7.3). AES narrowing was found in some participants but not all.

Thus, in nailing down physiologic correlates of the brightness given that the pitches and the register quality were the same for all singers the researchers couldn't know without actually looking at the video pictures because everybody did it in a different way. P10 found it tricky to decide on the exact mechanism that was responsible for the brightness of the sound in Track 20. He suggested that either a higher larynx or AES or a raised tongue were involved (see his quote in 6.7.3).

Our participants gave very different responses to the same question: "How do you make a dark sound?" P12 stressed short, thick folds, pulling everything back and dropping everything down. P11 described bright vowels as "smily" and "wide" and related dark vowels with a narrowing inside. P02 suggested that the brightness is controlled by the position of the tongue (see their quotes in 6.7.3).

Another physiological component is mentioned in relation to brightness: the middle constrictor. In P10's view it has an acoustic response similar but slightly less than the AES. P12, in turn, claims that the role of the middle constrictor is greater in creating brightness than AES.

We look in more detail at brightness in Section 6.7.3. As mentioned above, these are complicated, understudied and not exactly defined phenomena. P11's study though provides justification to assume that one of the reasons for such a discrepancy is that singers have different strategies to produce bright or dark sound. P12 confirms the difficulty of an explicit description, saying that he would give a very low confidence value for his ratings of vocal tract physiology responsible for brightness (see his quote in 6.7.3).

Thinning out the vocal folds

Now after we investigated different ways to achieve the same acoustic result, let us turn to a physiological setting – thinning out the vocal folds. Even here, our experts give us examples of different ways to achieve it. For Jo Estill vocal folds are thinned by tilting the thyroid cartilage. Yet P15 notes that there are more than one ways to thin the vocal folds (see her quote in 6.5.1).

P14 stresses that vocal folds are thinned in different registers, that they are "stretched long and thin in both head voice and falsetto, but in head voice a larger part of the vocal folds vibrates, allowing for more vibrato." P12 adds a dimension to Estill's view, emphasising that vocal folds' stiffness plays an important role and can counteract the acoustic changes produced by stretching and thinning them (see his quotes in 6.4.3).

The same holds for thickening the vocal folds. You can adjust the position of the crico-thyroid visor, e.g. straighten the thyroid cartilage. Whether cricoid cartilage can be tilted independently is speculative. What matters here is the relative position of the cricoid to the thyroid as P12 stressed. Alternatively, vocal folds can be stiffened as above "to make a massive sound" because a larger mass of the folds vibrates. The singer can also employ aerodynamic non-linearity trough narrowing the epilarynx – this will according to P10 make the singer feel and sound as if his vocal folds were thicker than they actually are.

7.5 Language and phonetics

> "What is the difference between singing and speech? In singing the vowels are longer, that's all." (P08)

Language and speech were mentioned by our experts as important factors in analysing singing quite often. There is an ongoing debate on the evolutionary origins of both speech and singing, on how they are related and whether singing preceded speech in humans or vice versa. We are not in the position to contribute to this debate here, but, as several of our experts stressed, the phonetics of the sung words, the singer's habitual articulation, the specificities of the spoken dialect or language play a crucial role in vocal production.

Speech position

A number of concepts mentioned by participants are related to speech. For example *speech range*, which is the range of pitches people comfortably use for speaking; P01 describes a vocal production at the limit of this range as "going into urgency":

> "Speech range, but going into urgency, not relaxed. Rather thick fold, probably on the limit where thick folds switch to thin." (P01)

Speech position is a widely used term, usually implying thick vocal folds or chest register. Jo Estill describes *speech mode* as a neutral vocal production with thick vocal folds and no thyroid tilt, with false vocal folds neither constricted no retracted. At least six of our participants used the terms *speech position*, *speech mode*, *speech-like sound* or similar. Some of them had a more specific understanding of speech position. P08 explicitly equaled speech position with vowel shape – she teaches her students to retain the formation of the vowels they use in speech for their singing. According to P08, while some musical styles, notably Western classical singing, require changing the shape of the vowels, in the traditional style she teaches keeping the *speech position* (the way to shape the vowels) is essential, and in most styles it supports efficient vocal production, clear diction and ability to express emotions in a natural way. There is a tendency in some people (particularly with Western musical background) to shape the vowels farther back when singing, therefore it's essential to counteract this tendency through vocal training. She expects from her students to "sing with their own voice" – this technical term means keeping the shape of the vowels the same as this particular singer would employ in their own speech.

She arrived at this concept at the beginning of her teaching carrier to support her students in learning the local traditional singing style. When she came across the concept of *Speech Level Singing* by Seth Riggs (1992), one of the most influential US American vocal coaches, she felt reassured in her approach.

P08 also discussed the role of the soft palate in retaining the speech mode at the passagio break and here she is backed by P15 (see Section 6.9.1). P08 suggested that the anterior part of the velum needs to be relaxed when approaching the passagio in order to stay on the speech side of it – here apparently referring to thick folds/chest register as speech voice. In P08's opinion spanning of the soft palate sail is responsible for the passagio actually taking place and vocal production changing from chest to head register.

P14 in turn mentioned speech falsetto, as opposed to head voice – vocal production only employed by trained Western opera singers according to P14. For him chest register was the configuration of the vocal apparatus typical for speech, but apparently falsetto was also related to speech.

To summarise, speech mode and speech position, while being one of those widely used terms usually taken for granted, can refer to a whole range of physiological aspects, including thick vocal folds, a general neutral configuration of the vocal apparatus, vowel shape and placement, soft palate activity and more.

Articulation, vowel shape and granularity of analysis

P08 described singing as opposed to speech as protracted sounds, vowels (see quote in7.5). P11 insisted on looking at sustained vowels as the basis for the analysis of vocal production (quote in 6.9.8). P10 stressed that vowel shape absolutely impacts on what's happening in the larynx (quote in 6.5.2). P12 noticed that you would have to modify physiological configuration with every vowel change (quote in 6.8.2). And P01 added that these changes affected her perception of singing differently for each vowel.

These are strong claims that contradict our approach. In Section 4.4.2 of the Methodology chapter we discussed the choice of time scale for our study and justified it by two considerations: a) the entities of analysis should be small enough to be physiologically stable and to provide a more detailed analysis than in Cantometrics, and b) they should be large enough for results to be comparable to the Cantometrics ratings.

We constructed our ontology for physiological analysis on the scale of several

seconds up to a minute and even more: while musical samples were between 30 seconds and two minutes, the snippets – the entities of analysis – were 6.7 seconds on average (with one outlier of 53 seconds). The length of the samples (about one minute) is enough for human listeners to rate perceptual characteristics of the songs. Each sample contained one or two snippets with differing physiological settings. In this situation one or two entities of analysis gave the listener an idea of physiological configurations used in the whole sample, and allowed to potentially compare physiology and perceptual descriptors.

Given the importance of vowel shape for vocal analysis it raises a legitimate question about the validity of our approach.

An average speaking rate in English is about 4 syllables per second (Cruttenden 2014), and this seems to hold for other languages (Osser and Peng 1964, Barik 1977). While vowel change in singing would be slower, it would still be on a tenths of second scale. To analyse each vowel change, one would have to deal with up to 240 entities of analysis on average to understand a one minute snippet. This is impractical if done manually and would not facilitate comparability with descriptors on a larger scale. For this reason we excluded physiological parameters which tend to change with each vowel change, such as jaw opening and lip form.

One of our participants implicitly supported our view that such an approach was legitimate. She compared two snippets from the same sample, saying that they were "physiologically the same", the only change being a different vowel at the lower note giving it a different timbre (see quote in 6.9.8). She implied that normally she would discard the vowel related jaw-lip-tongue information in her analysis; it was the large timbral difference that made the change relevant in this case. Three of 13 participants mentioned the jaw as an important factor in their interviews, and only one mentioned mouth corners.

Our musical examples were deliberately chosen so that there were not too many physiological changes in them, we have excluded all tracks with lots of change. A good example of singing where physiology changes frequently is yodelling. Therefore in some cases we cannot avoid looking at a large number of snippets to analyse.

At times our experts did find that articulation, vowel shape was very significant for their analysis (see Section 6.9.8 with many examples). Sometimes they incorporated vowel analysis into our suggested scheme, e.g. when the changes happen every time a given vowel shape occurs. For example P01 heard consistent changes in track 18 every time the A vowel was sung (see her quote in 6.9.2).

We included the position of the tongue in the ratings to test whether our par-

ticipants would be able to identify and describe its position and contribution to the sound on the macro level. Our interviewees used quite a varied vocabulary to talk about the tongue, and in many cases they said that the position of the tongue was not very relevant for the overall physiological setting. Often the tongue only became relevant when it inhibited other aspects of vocal production (see quotes in 6.8.2 and 6.9.7). We hoped that better vocabulary to describe the actions of the tongue would emerge in the interviews. Unfortunately that did not happen. There was also no agreement on the ratings of the suggested categories for the tongue. It might be worth considering to discard it in future studies.

It seems that overall our experts found the scale of analysis we chose comfortable and practical, not limiting their analysis but facilitating it:

> "I found it super, because at the first listen I didn't notice it (the change at the "a" vowel). Only on the second snippet, where you singled it out, was it clear that there is a change. It was great that you separated it." (P01, Track 18)

From speech to singing

> "Very difficult to say if you haven't heard that person speaking and what their total range is." (P15)

We often heard from our experts that having no access to the singer when analysing singing had a significant impact on the result. Seeing the singer gives the rater an idea of his size and build, as well as visual information about how their body acts and is affected by singing. Another source of understanding vocal production is hearing the same voice in a different context – at a different pitch, volume, tempo, vowel. The most important context other than singing is speech. Hearing someone speaking gives the expert a good idea of their individual traits and habits which can originate from their physical built, their experiences and their culture.

> "He must have a very high speaking voice, I think, because otherwise you couldn't sing like that." (P14 on the singer in Track 30)

Often in traditional singing (where no vocal training with timbral or phonetic standards is involved) singers employ the same vocal mechanisms in singing as in speech. One of the most important and obvious is shaping the vowels. P08 teaches traditional singing in an Eastern European country and pays a lot of attention to the

dialect which is sung, because the shaping of the vowels in the dialect has an immediate effect on singing. Individual characteristics of the singer and his speech are also as important. P08 stressed that we should take phonetics - the shaping of the sounds - into account because it affects vocal timbre (see her quotes in 6.9.8).

The velum is mentioned several times in relation to its rating depending on the speech patterns (see quotes in 6.8.1). And it's not just articulation, the source can also play a part. P08 gives an examples of a "slightly pressed sound":

> "this might be a speech pressedness, not vocal, because this is how he speaks. Some people will be speaking like that." (P08)

An interesting conversation took place in the interview with P01 on the relationship between singing and speaking pitch, while the participant was rating Track 24 from Taiwan. The female singer sings quite high and thick, and P01 was wandering how she speaks, whether her pitch would be similarly high in speech. Maybe women speak on a higher pitch in her language? Then her habitus would already be very different. P01 is bilingual speaking English and German, she noted that there is a considerable divergence between these two languages in terms of pitch, her English speech employing more head voice and a larger range. The interviewer noticed similar differences between her Russian and her German speech, Russian being still lower in the chest than German. They then contemplated on how these differences come to be. Do you possess a higher breathing (neutral) position of the larynx when you are born or grow up there? Is it inborn or taught, and to what extent?

The next quote sums up the relationship between language and singing and brings us naturally to the subject of our next section – familiarity with the tradition:

> "All of our impressions of voices are actually culturally located as well as physiologically located. Given different linguistic patterns I'm going to have different sorts of capabilities. And that's something I found out when I've been teaching and traveling. What we previously thought about the voice is actually very very much based on Western lyric vocal production and Western speech patterns." (P15)

7.6 Familiarity with the culture

When you decipher physiological processes behind vocal sounds the result will obviously depend on where you stand: your own physiological "default values", physiological dimensions in which you can perceive change and the span of these possible

changes, in other words, your mental representation of the vocal physiology space. In particular the ratings our participants provided in their interviews were based on their perceived vocal physiology spaces. P11 explained this fact with an example:

> "if you say to me as someone who was trained as a classical singer to evaluate what I've just heard against what I would know of Western classical singing I would evaluate it very differently than if you gave me an example of somebody's singing in the United States, country music from Appalachia, which would be very similar. So the context in which the evaluation takes place and how one quantifies the degree to which the physiologic function is extreme is dependant upon a lot of other things." (P11, Track 24)

It seems reasonable to assume, and it is confirmed by our interviews, that these perceived physiological spaces are to a large extent determined culturally: reflecting what kinds of sounds the rater is used to hear, to sing and to analyse. Our interviews point to the conclusion that if you are not familiar with the sounds made in a particular culture, if these sounds are not part of your musical vocabulary, if you cannot recreate them yourself you probably will have difficulties deconstructing physiological processes behind those sounds reliably. There were several examples of our participants trained and educated in Western classical music having difficulties rating musical examples with unfamiliar vocal sounds. There was also an example where a participant teaching in a non-Western culture ran into similar difficulties when rating singing from an unfamiliar tradition.

The interviews point us to what really makes a difference to the ability to deduce physiology: when the rater can recreate the sound easily he is most confident in his ratings. It must be noted that recreating the sounds does not guarantee that the rater's judgement is correct, due to the fact that similar acoustic results can be achieved through different physiological settings, as we discussed in Section 7.4. But the inverse conclusion is supported by our interviews: if you can not easily reproduce the singing you analyse you would probably have difficulties analysing it.

Some of our experts reflected carefully on the process of analysis and their perceptions. We discuss the controversial subject of empathic listening below, which is a common instrument of analysing physiology. While it provides raters with direct perceptions of what happens in the body of the singer, it can go very wrong if the rater is not familiar with the sounds or rater's voice is quite different from the singer's. And because this is a kinaesthetic perception, it is very hard to resist and

requires a careful self-reflection.

While the fact that our ratings depend on familiarity with the musical tradition seems obvious, the extent of misjudgement in the analysis and of possible bias in rating is not, and the consequences in particular for teaching singing can be very serious. As singing teachers in the UK and the US note (Kayes 2013, p. 5, Weekly and LoVetri 2009) it is very common at Western colleges and conservatoires that singing teachers have classical background but teach other styles such as music theatre or contemporary commercial music, which demonstrates institutional bias towards classical singing and is detrimental to the quality of teaching in non-classical styles. Our study broadens this discussion to include teaching vocal production beyond the scope of Western styles. Our findings support the argument that being able to produce particular vocal sounds with ease is crucial to the ability to judge the physiological processes in the student and therefore to coach their practice correctly.

From the Western point of view

For our participants who were only at home with Western classical music our ratings constituted a real challenge. For example P14 regarded chest voice as a configuration typical for speech and head voice only to be employed by trained Western opera singers. Head voice for P14 is characterised by a tilted spectrum with a high fundamental, as opposed to non-opera singers in chest voice, where the fundamental is low and the spectrum is dominated by other overtones (see his quote in 6.4.2).

While P14 was struggling to apply the classical registration terminology to examples of non-Western-classical singing, he was even less comfortable with Estill's thick vs thin vocal folds terminology. This led to some bizarre ratings, such as in Track 30, where he chose thick folds because the voice sounded speech-like, but stressed that the singer's vocal folds are thin and stretched. He declined to rate this voice as a mixture of thin/thick or chest/head, because, in his view, to be able to produce a mixture one had to be proficient in head voice, and this singer, not being a Western classical singer, could not be:

> "Head voice would be, the vocal folds would be stretched, like in falsetto, but they would vibrate with a better closure, and you'll see the vibrations more clearly then in falsetto, where the vocal folds just barely touch each other. So that would be something that opera singers would do in a high tessitura. But this is not an opera singer. So he would mix between the chest register and sometimes going into falsetto, more

speech mode." (P14)

In summary, for P14 head voice meant a very specific phenomenon which is only characteristic of Western classical singing and is acquired through training. He struggled to apply the classical registration terminology that was originally developed for Western classical singing in a different, wider context. He did not expect to hear head voice in our musical examples, because those singers were not trained in Western classical singing. He was therefore forced to shovel most of singing in our examples into chest voice or falsetto categories. But then, when he heard vocal production that did not fall into these two categories (Track 22), he resorted to using head voice and ran into a contradiction.

Another participant, P02, a European phoniator and a senior staff at a hospital, himself a classical singer, admitted that he only recently had come to visualising physiological settings not related to Western classical tradition:

> "Concerning open and closed sound, this feeling of narrowing, I have only come to realise after Johan Sundberg's course, because this is something one is not aware of as a classical singer." (P02)

P12 mentioned microtonality which is not normally present in Western music:

> "I'm not used to micro tonality, which is part of it. That doesn't mean it is wrong. We use it to experiment with Western traditions. It's interesting." (P12)

As a voice professional you have to make judgements about the health of someone's voice and about vocal efficiency. Tense vocalisation is often synonymous with unhealthy and inefficient. But what is considered unhealthy and inefficient in Western vocal space does not necessarily violate vocal preferences in another culture (see quotes in 6.1).

We discussed the crucial role of language phonetics for vocal production in Section 7.5. Apart from vocal health and efficiency it is often the aesthetic quality of the singing that we form our initial judgement of, which then influences the further analysis (see P10's quote in 6.3.2), and this judgement is highly culturally biased.

From personal experience

A personal experience with the sounds to be analysed is the best position for the rater to form confident and reliable judgements.

P11 who sings and teaches in various contemporary commercial Western singing styles found it easier to relate to the sounds she heard in our examples. She could imitate the singing in Track 24 without strain. But she stressed that Western classical singers would find it difficult and would therefore reject it:

> "The freedom of that sound: [sings] This sound I'm right now at this moment making for you, I can do with ease, with great ease, I do not feel any strain. And so I could, if I were a person who was going to be in that culture, I could easily sing there and be very happy. I don't know if that would be something that would allow me to sing in other styles, but the comfort range of the person singing I would say is quite good. In comparison to someone who is from the West who've been trying to sing classically, who would try to make that sound who would be feeling like they were dying! Oh, I can't make that sound, the throat hurts, oh, that's so awful! " (P11)

Another example of misjudgement due to unfamiliarity with the musical culture was the analysis of Track 18 – a Gamelan piece from Java – by P08 who teaches singing in an Eastern European tradition. For her the vocals in the snippets sounded very high and her ratings were based on this notion. Until the moment when she checked the actual pitch which turned out to be an octave lower than she expected. What tricked her was the singer's very high position of the larynx. P08 had to revise all the previous analysis in view of this new information. In contrast, P01 was aware of this contradiction from the outset (see her quote in 6.4.4). P01 had personal experience in singing Indian music, where the vocal production technique is similar to the one the singer uses in the given example:

> "You should know, I have worked with Indian music, therefore I am familiar with sounds with very high larynx and twang." (P01)

She worked with Indian musicians and learnt to sing this way in workshops. This is how she describes her experience in a Dhrupad singing workshop:

> "You go down, and down another bit. You sing in a sitting position and try to attain the most relaxed posture you can. The teacher worked with Feldenkrais and Dhrupad singing, and we did lots of relaxation exercises, and then checked the voice – how is it now? I think the main point was that you approach it with so much less tension. ... The crucial thing to be able to sing this way was the low breath pressure." (P01)

With this experience it was obvious to the participant that she would perceive the singing in the example as very relaxed. She was also fairly confident in her ratings on examples 18 and 24 with this kind of vocal production.

Yet in another interview episode, when discussing Track 26, P01 had difficulties to reproduce the lower notes where the singer kept her larynx high. She perceived the singing as tense, pressed. Aware that this perception might come from her own limitations she tried to sing lower notes with a high larynx and commented on her difficulties (quote in7.6). At the same time, the interviewer could easily reproduce this kind of vocalisation. P01 mentioned that she worked with her students on the ability to keep the larynx high when going down in pitch. At the same time the fact that she could not easily produce these sounds herself led to difficulties when deconstructing and rating physiology behind the sound.

This example leads us straight to our next point about what channels we use to perceive and analyse physiology.

Empathic listening

In our study auditory information was the only available element for our participants to analyse. Yet they often spoke about how the sound "feels". In particular, one participant was very reflective on the process of analysis and apart from listening she emphasised what she called *empathic listening*:

> "I always listen empathically. I try to feel into it. I've just tried to feel whether the pressure remains but the airflow decreases. I'd say it is low." (P01, Track 16)

She refers here to the phenomenon of rapport or motor imitation, when the body of the listener quietly and usually unconsciously guesses and reproduces the physiological process of vocal production that takes place in the body of the singer.

When analysing Track 26 where she had difficulties with the high larynx on lower notes, P01 reflected carefully on her perception of the sound as pressed: she wondered whether this impression appeared because the singer changed her physiological setting (and she as rater could deduce it from the sound) or whether it came through empathic listening:

> "I've tried to filter whether it is the sound itself that I perceive as pressed or whether she brings in something [a new physiological setting that increases pressure], because in principle I don't expect the voice to sound

> pressed here. Maybe it's just me, since I listen very empathically, that my larynx resists when I listen to this, and therefore I feel it that way." (P01)

This phenomenon has been researched and debated for decades in linguistics, under the auspices of motor theory of speech perception (Galantucci, Fowler and Turvey 2006). Its main hypothesis states that listeners perceive speech by identifying the vocal tract gestures with which they are pronounced rather than the sound patterns. It was shown that hearing speech activates vocal tract muscles (Fadiga et al. 2002), the motor cortex and premotor cortex (Watkins, Strafella and Paus 2003, Wilson et al. 2004). There is evidence that perception and production are coupled in the motor system supported by the existence of mirror neurones, which are activated both by hearing (seeing) an action and by carrying the action out (Rizzolatti and Craighero 2004).

From our experience of analysing singing and from what our participants said we conclude that in most cases a vocal production analysis will be a mixture of both acoustic analysis and empathic listening. Both our hearing and our kinaesthetic reactions are subjective. Analysing acoustic events, if done by experts, can be more objective, yet it is only as good as our knowledge of vocal physiology models. And as mentioned before, this is still limited, in many cases there are multiple ways to interpret the same sound physiologically or no plausible interpretation at all. While empathic listening gives us a tool to overcome these limitations and access bodily information directly by means of rapport and intuition, these unconscious reactions are shaped by the actual experience of the analyst. If the mechanisms used by the singer are not familiar to the analyst, their body would most probably switch to physiological processes that are more habitual trying to recreate the sound. This is how it comes that a singer sings comfortably while the listener's throat suffers the most painful effort.

> "when they listen, and then go: "Oh, that hurts my throat when I listen to that!", I think well, but it doesn't seem to be hurting the throat of the person doing the singing, does it?" (P11)

Another barrier is the individual difference between voices, in this case between the singer and the analyst: what is easy for the former may be harder for the latter just because their vocal apparatuses are different:

> "there is a thing called projection where we project our experiences onto other people. And that's a problem with singing because what's easy

> for me may not be easy for you and vice versa. But if we teach from the standpoint of everything that is easy for me should be easy for you then we get into trouble." (P11)

Empathic listening is subjective and can play tricks with us when we rely on it for vocal production analysis. Being a kinaesthetic reaction it is very hard to resist the impressions we get from it. It is therefore important to be aware of its subjectivity and limitations and it requires a careful reflection, like the one we see in P01's interview:

> "You mean when I come from my empathic listening? I feel that the area around the vocal folds is open." (P01, Track 16)

And especially when we are faced with sounds and vocal techniques we are not familiar with, we run into danger of reinterpreting what we hear in familiar terms or getting our physiological setting into conflict with itself. It is our own ability to recreate the sound that would give us a certainty that we would not experience an excessive level of effort when listening to unfamiliar sounds and thus grossly misinterpret what happens in the singer's body. Tracks 18 and 24 are excellent examples making the importance of personal experience clear: while P08 misinterprets the pitch due to a high larynx position (it feels high, so it must be a high pitch), P01 is aware of this configuration because she has experience singing such music. P01 also mentions the velum as a source of contradiction, again referring to the kinaesthetic perception (see her quote in 6.8.1).

Tradition vs physiology

The broad question of how tradition or the enclosing culture influences singing was not part of the questionnaire, but it was nevertheless addressed by some of the participants. We have discussed the influence of language and phonetics on singing in detail in Section 6.9.8. For example, a possibility was considered that the neutral larynx position varies across languages and therefore will be different in singing. In this way cultural preferences, here – phonetic habits – can directly influence physiology in singing. Another source of cultural dimensions or, depending on the angle, limitations come from local music styles, performance practice and preferred vocal production. Like one's language and phonetics are formed by their environment, in the same way one's vocal production is to a large extent dictated by what is customary around them, what they hear and consider as right or preferable.

P15 mentioned both language and the surrounding musical tradition as important factors when evaluating efficiency and tension (see her quote in 6.1).

If you hear an elaborate Muezzin call five times a day, you would inevitably try to recreate the sound in your singing (as long as your voice has the capacity for it). If you audition for musical theatre today, you'll have to be belting up to C5 and you probably have been developing this demanding technique either intuitively or through formal training. And if your vocal apparatus is not well suited for this kind of sounds due to individual characteristics you most probably would not consider yourself as a singer (at least not a musical theatre singer) or the society would not confirm you as such. Thus, apart from a direct impact on physiological settings, the surrounding culture serves as a double filter: firstly, physiological mechanisms relevant for culturally preferred vocal production are developed; secondly, singers with relevant vocal abilities are given preference above all others. P08 contemplates:

> "Why did this particular kind of sound form, how is it related to the tradition, why such and such requirements are placed on, say, female singing? Like Japanese women who had to wear very small shoes so that their feet became smaller, same here, traditional preferences are imposed on the larynx position, on the vocal folds length if you want. If you always sing with this kind of sound, they change, their habitual stretch changes. It would probably change the speech as well." (P08)

Another example of such cultural differences in vocal production was the discussion on the larynx position on the lower notes. While P01 found it natural to lower the larynx with lowering pitch, which is a norm in many Western musical styles, for the interviewer it was very easy to keep the larynx in a high position while moving down in range, because in her singing tradition it is a common technique.

To be precise, when we talk about tradition affecting larynx position or the stretch of the vocal folds, we are not assuming that the basic human physiology changes in any way. What is determined by these cultural preferences are vocal habits, the kind of sounds (and physiological settings) which are considered as a default, as easy or natural. P11 for example differentiates between habitual and unconscious constriction on the one hand and contraction done for musical purposes which is deliberate and controlled. She relates the former type with tradition: if someone makes a sound that is typical for them and is the only sound they know, even if she perceives it as constricted, it is opposed to tightening something on purpose (see her quote in 6.7.4).

Being an anthropologist Alan Lomax was intrigued by the relationship between singing and tradition in a broader way. He came to the conclusion that societal mechanisms – patterns of communication established in a given society – should be governing all aspects of life and coexistence, including performance practice. Therefore, singing, which is a highly regulated communal activity, was as good a reflection of these patterns as anything else and a society could be studied on the basis of this reflection. He was looking for a musical core of a society, convinced that crucial aspects of performance practice would manifest themselves in almost any musical utterance from the given culture. He also found that in the process of compiling his dataset ten musical examples from a culture saturated the variety in ratings, that no new musical descriptors were added to a culture with a larger than 10 set of musical samples. This was a supporting evidence for him confirming his hypothesis about musical core. The geographical clustering of musical profiles of cultures was another.

The idea of musical core was severely criticised by ethnomusicologists, who brought numerous examples of a wider variation in performance practices than could be represented in the Cantometrics dataset. It remains open though whether it was the limitation of Lomax's musical sources that caused profiles to converge for sample sizes larger than 10 or whether his theoretical assumptions were flawed. It seems plausible to assume that communication in an orchestra is governed by the same laws as any other communication in society; it is much less obvious that larynx position and nasality could also be subject to these laws. To investigate the latter assumption has been an important motivation behind our current work: automating the analysis of singing would allow to scale the dataset up from ten samples per culture to any size, ultimately to include all recorded music.

8 Discussion and future research: Cantometrics, MIR, singing education

We began our research for this thesis motivated by the idea to revisit the Cantometrics experiment and to revise its approach using more objective descriptors of singing and contemporary computational techniques. Cantometrics addressed the core question of ethnomusicology – the relationship between music/singing and society, using the methodology of a large-scale statistical comparison, routinely used in many disciplines but rejected by ethnomusicologists. Cantometrics findings are stunning and controversial, and have rightly been assessed and criticised within the field. But even more than the findings themselves, the question of the validity of cross-cultural comparative approach remains divisive. The majority seems to be wary of wide comparisons, and Cantometrics is sometimes seen as the case to prove it. Yet we are still to see a systematic analysis of where and why the experiment failed. Even less attention has been paid to what might have been valuable and what we could learn from Cantometrics. Using contemporary technology we may potentially overcome many methodological weaknesses, such as a limited number of samples per culture, unrepresentative data, etc. Our thesis aimed to close this gap, to present a detailed analysis of one descriptor, its reformulation in more objective, better measurable terms and to advance the Cantometrics methodology with the means of machine learning and automatic classification. While we fell short of that ambitious goal, we have paid particular attention to what underlying reasons and confounding issues hindered the progress and which further improvements are possible given the current state of knowledge and technology.

We developed an ontology of vocal production consisting of 17 descriptors / 29 dimensions comprised of objective descriptors which refer to vocal source aerodynamics and vocal tract physiology (Chapter 2, Table 2.2, see also Table 6.2 for an updated ontology). We then investigated two different approaches to the complex

task of revising Cantometrics.

For the incremental approach an initial proof-of-concept experiment on automatic labelling of phonation modes for sustained sung vowels was presented, followed by a discussion of generalisation to more complex data (Chapter 3).

The integrated approach encompassed all the variability of the original data. We developed a method to annotate the Cantometrics recordings by means of expert listeners knowledge elicitation (Chapter 4). In an investigative mixed-method study we interviewed 13 experts collecting quantitative and qualitative data. A tendency to inter-rater agreement was found for descriptors *AES* and *larynx height*. No agreement was detected for other nine investigated descriptors. (Chapter 5).

For the agreed parameters *AES* and *larynx height* we collected the average ratings – reliable annotations of vocal production given the current state of knowledge (Table 5.7). The 19 snippets from our musical examples together with *larynx height* and *AES* annotations constitute the first ever cross-cultural dataset with reliable annotations on vocal production, published and curated at the Open Science Framework[1]

We confirmed our original hypothesis that we can map the Cantometrics *vocal width* onto more objective descriptors of vocal production - larynx height displayed a strong correlation to *vocal width* (Section 5.8). This finding can be used to further confirm or refute the relationship with the subordination of women.

It was surprising though to find no agreement between our experts on the majority of our descriptors. As with any negative result, its value is in the analysis of problem cases and confounding issues, which we think is an important contribution of this thesis (Chapter 7). In this chapter we want to summarise and discuss what we learnt from the perspective of three research fields: Cantometrics, MIR and singing education. Each section contains a summary of our findings and their implications directly relevant to the respective field. Additionally, each section will address revising the Cantometrics experiment from the point of view of the respective field: we suggest a list of steps necessary to achieve that goal in Section 8.1.6.5 and discuss it with regards to MIR (Section 8.2.4) and vocology/vocal pedagogy (Section 8.3.3). We highlight the need for collaboration between the various fields to address a large-scale project like Cantometrics in the contemporary technological reality and provide a roadmap for such a collaboration. Further suggestions for future research not related to Cantometrics are listed in Sections 8.3.3 and 8.2.5.

One of the contributions of this thesis is opening up a dialog between research

[1] `https://osf.io/pff8m/`

fields on vocal production. In addressing a complex problem such as Cantometrics revision, which can potentially have far-reaching implications in a range of knowledge domains, we establish a new line of thinking expanding beyond the limits of each single field involved: ethnomusicology, MIR, vocal physiology, voice science, etc. Due to a combination of expertise this work presented advances in each of the fields, while taking the interdisciplinary research question a significant step forward, presenting a roadmap for further collaboration.

8.1 For Cantometrics

8.1.1 Vocal width/vocal tension

In this thesis we compiled a formal vocabulary to describe vocal tension that allowed us to deconstruct the notion of *vocal width* introduced in Cantometrics (Section 2.3). We suggested that the three aspects mentioned in the Cantometrics definition - narrowness, tension and resonance - are not directly related and hypothesised about a possible contribution of physiological descriptors from our ontology to the perception of *vocal width/vocal tension* (Section 2.4). Our interviewees confirmed the absence of direct relationship between tension and narrowness (see analysis of Track 24 in Section 6.1) and generated numerous discussions about various descriptors relevant for their perception (Chapter 6). We update our suggestions for the possible contribution of physiological building blocks to *vocal width* in Section 8.1.1.2.

We also found two descriptors for which our experts provided consistent annotations - *larynx height* and *AES* - whose ratings correlate strongly with *vocal width* ratings for the eleven tracks we investigated (Section 5.8).

8.1.1.1 *AES, Larynx height*

AES and *larynx height* are the two descriptors which were found to be consistently rated by experts and to correlate significantly with *vocal tension*. Therefore they can be used as a more objective alternative to Cantometrics *vocal width*. We would expect the correlation between *vocal width* and pre-marital sex sanctions on women, discovered in Cantometrics, to be retained for both *AES* and *larynx height*. These physiological descriptors can be rated for new tracks by expert listeners in a procedure similar to ours. This approach would be advantageous in comparison to rating *vocal width*, which would require a proper diversification of raters.

Yet it is important to remember that both *larynx height* and *AES*, although they

were rated consistently in our study, lack a precise and unambiguous definition. While the anatomical and physiological definition of the aryepiglottic sphincter is unambiguous, its function in singing certainly isn't. Its activation is associated with twang and belt in the Estill model (see Section 2.1.3.2). Yet this notion has been contested by a number of our participants, with one vehemently stressing that vocal source is the primary mechanism where belt is generated (Section 6.7.1); with other experts claiming that this acoustic result can be achieved by multiple physiological strategies, which may or may not involve the aryepiglottic sphincter (Section 7.2).

Another source of confusion was the fact that a number of our participants were not familiar with the term *AES*, neither in its physiological nor its functional meaning. As one participant claimed, *AES* is a purely Estillian term, which no singing teachers outside this community ever use. Therefore the participants with no knowledge of this term were introduced to it via other descriptions such as *twang*, *small space* and *bright sound* (Section 7.1). We discuss the interplay between all these terms, their functions and physiology in detail in Section 6.7. We arrived at the term *narrowness of the supraglottic vocal tract* as a common denominator (Section 6.7.5). It is even more astonishing that with all these controversies and ambiguities our participants agreed about the values of this particular descriptor.

Larynx height or position of the larynx is measured in comparison with the resting position in speech (higher or lower), but in singing it is affected by the pitch as well as the register or the mechanism: for the chest register, the higher the sung note, the thinner the vocal folds have to be, the higher the larynx moves to stretch them. Apart from producing the required pitch singers are capable of moving the larynx higher or lower of the position most comfortable for the given pitch production to change vocal quality and colour, to express emotion (Section 2.1.3.2). As we have seen in our experiment the movement of the larynx is recognised most reliably by our experts with a good agreement about its position even in unfamiliar musical cultures (Section 6.6).

8.1.1.2 Physiological descriptors contributing to Cantometrics *vocal width* – an update

Our statistical results for *larynx height* and *AES* are significant in spite of ambiguous definitions of the terms. Interestingly, the term for *AES* that found the widest support – *narrowness of the supraglottic vocal tract* – literally refers to one of the dimensions of *vocal width* – the narrowness. Additionally, and somewhat

controversially, experts mentioned the tongue and the jaw opening in relation to the narrowness of the supraglottic vocal tract, which puts it even closer to the Cantometrics notion of *vocal width*. Yet nowhere in the discussion was tension or "squeezing" ever mentioned. It is also telling that *vocal width* correlation to *AES* was weaker than to *larynx height*.

For other ontological descriptors, for which no inter-participant agreement was found, a statistical approach cannot be taken. Yet we can extract relevant information on their relationship to *vocal width* from our interviews. In Section 2.4 we discussed the hypothetical contribution of various aspects of vocal physiology to the perception of the three components of the *vocal width* parameter: width, tension, resonance. We can now update Table 2.4 with anecdotal evidence from our interviews – see Table 8.1.

The tongue is a large organ and its position in the mouth defines the shape of the mouth cavity resonator in vocal production. We therefore already marked the tongue as a contributor to width/narrowness. Position of the tongue is also known as one of the important physiological components affecting the brightness of the sound (see P10's quote in 6.7.3).

This quote also mentions *AES* and *larynx height* as main contributors to brightness of the sound. Brightness is a resonance characteristic, yet it is not well defined and the relationship to Lomax's opposition between "richly resonant" and "restricted in resonance" is not clear. We can only say that we most probably will see some effect of these physiological actors on resonance.

The velum opens and closes the nasal channel. Lowering the velum was named by P11 as one of the various techniques used by participants of her study to make their vocal tract smaller and achieve a brighter sound (quote in 6.7.3). In this respect a low velum could contribute to narrowness in Cantometrics. On the other hand, P01 mentioned a low velum as one of the characteristics of a deep relaxation needed for the vocal production in Track 18 (compare quotes in 6.8.1 and 7.6). Thus it could be a sign of relaxed singing.

Our last descriptor *position within register range* captures not only the registers used by the singer in the given vocalisation but also where within the range of this register for the given singer the sounds are: in the middle or below, high or very high (Section 2.1.2). Our initial assumption was that sounds that are very high within a register range would also sound tense and score more in the Cantometrics *vocal width*. We did not find any confirmation of this in our study, as well as no agreement between participants on the values of this descriptor.

Table 8.1: Ontology dimensions vs *vocal width* components – adjusted

ontology descriptors	**ontology dimensions**	**width**		**tension**		**resonance**	
		wide	narrow	relaxed	tense	resonant	restricted
phonation	phonation breathy			✓			✓
	phonation pressed				✓		
	phonation neutral						
	phonation flow	✓		✓		✓	
onset	aspirate			✓			
	glottal				✓		
false vocal folds	constricted				✓		
	retracted			✓			
AES	narrow		✓			?	?
	wide	✓					
larynx height	low	✓				?	?
	high				✓	?	?
velum	low		✓	?			✓
	high	✓					
tongue	low	✓				?	
	high		✓				?
	compressed					✓	

This semantic investigation could be continued on our data. Yet a more promising approach would be to examine the data on perception of width and tension we collected during the interviews and compare them with the responses on salience (see Section 8.1.1.3). That would give us a better insight into what physiology contributed to our participants' perception of width and tension.

8.1.1.3 Vocal width – next steps

Because our research was motivated by the Cantometrics *vocal width/vocal tension* parameter in particular we have collected a wealth of data that are relevant to its various aspects and are yet to be analysed. We have explicitly asked our participants to rate perceived dimensions of width/narrowness and tension/relaxation in the tracks before as well as after they performed the physiological analysis. This data can be used as an independent set of ratings of the Cantometrics parameter, con-

firming or refuting the good inter-rater agreement claimed by Alan Lomax (Ebel's inter-rater reliability coefficient of .92, Lomax 1977, p. 270). While Lomax's raters were ethnomusicology students, our sample of raters provides a significantly different demographic spread as well as a wider variety of backgrounds and musical experiences. The data we collected also gives us a unique insight into the effect of detailed musical/physiological analysis on the perception of vocal width.

Two other sets of data were collected in our experiment which are particularly relevant here: on confidence and on salience. The confidence ratings were provided by the participants along with each physiological rating documenting their level of confidence in this particular judgement. This was necessary because physiological settings can be very difficult to decipher. These values can be used as a weighted factors. Salience data is of a more qualitative nature: after analysing a track our experts were asked which descriptors were most relevant and least relevant for their perception of singing in the track as wide/narrow and as tense/relaxed. The responses were often very different between participants and depending on the track; sometimes interviewees found these questions difficult to answer. This unique dataset could provide meaning and knowledge based justification or otherwise to our purely statistically derived correlations between Cantometrics parameters and physiological descriptors.

We have shown that *larynx height* and *AES* were rated consistently on 11 tracks/19 snippets and that their ratings correlated strongly with the *vocal width* ratings. The next step would be to obtain more tracks with *vocal width* annotations and collect experts' ratings on *larynx height* and *AES* for them. That would allow for a further confirmation on a larger dataset of the strong correlation between *vocal width* on the one hand and *larynx height/AES* on the other. If this set is sufficiently large, machine learning models can be trained to automatically recognise *larynx height* and *AES* in recordings of singing. Then, if a corpus of tracks with anthropological ratings for subordination of women were available, the relationship between that trait and (automatically determined) *larynx height* and *AES* descriptors can be investigated. If such a relationship is detected it would confirm the Cantometrics findings on the relationship between *vocal width* and subordination of women, at the same time reformulating it in more objective terms. An absence of such relationship would refute the Cantometrics findings.

8.1.2 Other Cantometrics parameters: *nasality*, *volume*, *rasp*

While we were primarily interested in the Cantometrics parameter of *vocal width*, the richness of our physiological model allows all Cantometrics parameters describing vocal production to be expressed in terms of our ontology.

8.1.2.1 Nasality

The Cantometrics parameter of *nasality* is directly related to our *velum* descriptor: high velum means no nasality, mid velum corresponds to some nasality and low velum to full nasality in the Estill model (Yanagisawa, Kmucha and Estill 1990, Estill et al. 2005a, see Section 2.1.3.2). As with all Cantometrics parameters, Lomax claimed a decent inter-rater agreement on the values of *nasality* in his experiment (Lomax 1977, p. 270). We therefore expected to see a good agreement on *velum* at the outset of our study: it seemed to us that nasality can be more easily detected than many other characteristics and there had been enough evidence for its ratings to be straightforward. Yet the outcomes proved us to be wrong: there was no tendency to agreement about the *velum* descriptor. This is in contradiction to the Cantometrics results and deserves a more detailed study. It would be highly instructive to determine the differences in experiment design between our work and Cantometrics that could have led to the different outcomes. On a more general ontological level, recent studies have shed light on the polysemy of the term *nasality* among singing teachers, in particular highlighting language differences (Garnier et al. 2007b). There is even evidence that the link between velum opening and nasality is not as straightforward as was previously understood (Birch et al. 2002, see 7.1 for further discussion).

8.1.2.2 Volume

Volume is another Cantometrics parameter to be addressed. While it is a perceptual characteristic and is judged subjectively, it is related to the physical value of sound pressure level (SPL). Johan Sundberg showed that subglottal pressure directly affects SPL (Sundberg 1987): to increase the dynamics and to get a louder sound singers increase their subglottal pressure.

Along with increasing vocal source vibration amplitude directly by raising subglottal pressure, a number of resonance strategies are employed by singers allowing them to increase vocal intensity. The most widely known is the singer's formant – a clustering of F3, F4 and F5 formants that builds an energy peak in the spectrum

around 3 kHz (Sundberg 1987). This strategy is typical for male and alto singers in Western operatic tradition and allows them to carry the sound above that of the large orchestra without any further amplification. This technique is typical of Western operatic singing which is described by Estill as a combination of *Sob* (low larynx, ventricular folds retracted, intensive head/neck anchoring) and *Twang* (narrowed AES).

Another strategy that helps singers raise perceived loudness levels without increasing effort at the vocal folds is twang. Estill defines *Twang* as a combination of narrow AES, thin vocal folds, high larynx and a tilted thyroid (Estill et al. 2005b). In a more recent research Titze and Worley (2009) suggest that the resonance mechanism behind twang is an acoustic coupling between the glottis and the laryngeal tube, lowering phonation threshold pressure of the vocal folds and boosting sound levels in parts of the vocal range that may be less responsive to vocal tract resonance. This mechanism is facilitated by the narrowing of the epilarynx (e.g. via AES activation). From the vocal pedagogy side, Gillyanne Kayes also mentions twang as a mechanism contributing to loudness in CCM singing styles (Kayes 2013, p. 103).

These two strategies to raise the volume of vocalisation are distinct. For instance, Sundberg and Thalen (2010) found no clustering of F3, F4 and F5 in their study of 'twang' voice quality, confirming that twang is a different strategy to the singer's formant. Guzman (2015) studied how vocal intensity affects larynx position and found that in classical Western singing larynx tends to go down with increasing intensity while in twang larynx is raised. This confirms the distinctiveness of the two strategies and Estill's view of larynx position for vocal qualities Opera and Twang. Interestingly, in all of the above discussion we referred to the three descriptors – *AES*, *larynx height* and *subglottal pressure* – which showed a tendency to inter-rater agreement in our study. It would be worth investigating the explanatory power of the three descriptors for the original Cantometrics *volume* classification.

8.1.2.3 Rasp

Lomax does not give a definition of *rasp* apart from "vocal harshness", he deems it to be self-explaining and is more interested in its social correlates. Yet from physiological point of view harshness has many different faces: the voice can be breathy, creaky, irregular (yodely), hoarse etc. Harshness can result from a vocal health problem; it might be a reflection of singing or speech production habits; it can be a vocal effect used occasionally for expressivity or be a conscious choice of

the performer to acquire and use techniques producing various kinds of harshness more broadly – Louis Armstrong and Tom Waits are just two examples.

Vocal health professionals have learnt a great deal about vocal harshness and its measurement. This subject needs a separate consideration which we omit here; we just mention the most widespread measurement scales. GRBAS is a perceptual scale measuring Grade, Roughness, Breathiness, Asthenia, Strain (Hirano 1981). Jitter and shimmer are measures of the cycle-to-cycle variations of fundamental frequency and amplitude, respectively (Morris and Bernard Harmon 2010). Noise-to-harmonic ratio could be a good general spectral measure to determine a presence of rasp.

Phoneticians are also interested in physiological settings we consider as harshness because in some languages these may be used as phonological entities. Moisik and Esling propose a model of three levels of constriction: glottal, ventricular and pharyngeal – each of which affects the periodicity of vocal folds vibrations (Moisik and Esling 2011). Thus breathy voice implies incomplete glottal closure; it can result from disturbances on the surface of the folds such as nodules or polyps; it may indicate a less effective vocal mechanism like in an untrained singing voice or be a conscious choice of phonation e.g. to evoke sexiness (see Section 2.1.1.1). Vocal fry is considered to be a proper register/laryngeal mechanism where irregularity of vocal folds vibration is due to them being tightly compressed and thus slack, compact and heavy, with a loose closure, producing a characteristic popping sound when the air passes through them (see Section 2.1.2). Ventricular folds constriction may also intervene with vocal folds self-sustaining oscillation (Esling, Harris and Romero 2003). Growl as we know it from Louis Armstrong, also called epiglottic trill, is produced through co-vibration of vocal folds and aryepiglottic folds which results in subharmonic oscillation (Moisik, Esling and Crevier-Buchman 2010). Growl has also been studied in throat/overtone singing (Sakakibara et al. 2004a, see also our example in Section 7.3). Catherine Sadolin in her Complete Vocal Technique (2000) explicitly teaches her students to acquire and use various kinds of rasp (which she calls effects) safely in their performance.

We mention this multitude of different physiological settings that produce various kinds of rasp to stress the fuzziness of this Cantometrics parameter from the point of view of our current study. Given the wealth of approaches – phonetics, vocal health, singing education, spectral science, ethnomusicology – this parameter calls for another interdisciplinary study in the spirit of our work.

8.1.3 Was Lomax right? Objective vs perceptual evaluation of vocal production

The question needs to be posed here: Given that for most descriptors we found no inter-participant agreement, and Cantometrics raters did agree on their ratings, maybe Alan Lomax was right in his perceptual approach? Could it be that we can better agree on perceptual descriptors than on physiological ones which require expert knowledge, invasive measurements and are prone to fuzziness due to multiplicity of vocal strategies? Voice is after all a perceptual phenomenon in response to an acoustic stimulus. Why can't there be perceptual vocal characteristics that we could universally agree upon? Perceptual ratings of voice are subjective, but, as Oates 2009 notes, subjectivity alone is not a sufficient reason for rejecting this approach to voice assessment. Subjectivity does not necessarily mean that reliability and validity are inadequate. If intra- and inter-rater reliability and agreement were high and statistically significant, then these ratings would be preferable to physiological descriptors, in particular as long as there is no cheap and non-intrusive way to measure them objectively. One of our participants insisted that it is easier to teach any singer, even children, to recognise vocal quality, register and vowel quality (bright/dark) than to convey physiological processes (see her quote in 6.10).

This is very close to Lomax's idea of vocal parameters that are universally understood and can, after a short training, be rated by anyone.

In an analogy to our argument for or against perceptual vocal parameters (Cantometrics) vs objective descriptors (our study), Oates (2009) analyses pros and cons of auditory-perceptual evaluation of voice quality (vs objective measurements) for dysphonic voices in a clinical context. Oates concludes with the suggestion to use a mixture of perceptual and objective descriptors for best results in identifying dysfunctional voices. Following the parallel between vocal health and vocal production research we find that our participant P11 also includes both kinds of descriptors on her list: bright/dark or heavy/light can be highly subjective, while pitch and volume, though also perceptual terms as such, she relates to objective measurements of frequency and SPL. Garnier's (2007b) list of common terms in French classical jargon also contains both kinds of terms: bright, light, metallic, tight or efficient on the subjective perceptual side, but also reinforcing high harmonics, which can easily be confirmed through acoustic measurements.

In fact, while we use the objective language of vocal physiology in our current experiment, the mode of evaluation remains auditory-perceptual: we still employ

human listeners judging singing samples exclusively via listening, providing their subjective ratings. What makes our situation more defined than a general perceptual experiment is the fact that our participants are highly experienced in singing voice evaluation. Their prolonged experience justifies the assumption that their mental representations of singing which serve as their internal golden standard (Kreiman et al. 1993) have been acquired and tested on many instances over time and therefore do not change significantly within subject; we expect to see little intra-participant variation, for which we cannot control otherwise due to a limited interview time. Also, the fact that there is a physical reality behind descriptors, contributed, so we hope, to our participants' intention to be as objective as possible, in contrast to highly subjective and ambiguous parameters such as tension.

8.1.4 Why Lomax saw a better inter-participant agreement than we did

There is evidence that non-professional listeners converge to a consensus more easily in clinical evaluation of healthy and dysfunctional voices: Kreiman et al. 1993 notes that inter-participant agreement among highly experienced listeners is generally lower. This is related to the mental representations of vocal production which are much more nuanced and rich in experts. This may be one of the reasons why Lomax saw a better inter-participant agreement than we did. Another possible factor could have been that we introduced a wider diversification of raters – professionals from a variety of fields, from singing teachers to speech and language pathologists were involved, who would invariably have had very different internal representations of the voice. Yet interestingly we didn't find any professional bias in our data (Figure 5.6.2); either our sample was not large enough to see any effects, or other factors (possibly cultural and musical background, see Chapter 7) played a bigger role.

Another important feature of the Cantometrics experiment was the specific training on the parameters, in particular providing the raters with vocalisation examples of the scale extremes. This was necessary because their raters were not expected to have had experience in singing and most probably had never heard many instances of extreme vocalising. Providing them with examples ensured that the anchoring of the rating scales was the same for all raters (it was even tested after training with extra examples, the *Consensus Tapes*, see Lomax 1977). In our experiment, while we suggested the dimensions and terms for rating, we consciously avoided imposing boundaries on our expert participants; we were interested in eliciting their know-

ledge about vocal extremes. This approach made inter-rater agreement even harder to reach.

8.1.5 A word on musical universals

Musical universals have been a subject of interest in ethnomusicology for some time, with a recent upsurge in publications from interdisciplinary research groups. Our work offers a new angle on this area of research. In previous work the main research question was around the presence or absence of a given musical feature in all or in a majority of world's cultures. In fact, Savage et al. (2015) carefully account for historical/geographical dependencies to ensure that these factors do not bias the universality criterion. Yet neither their study nor, to our knowledge, any other previous work pay similar attention to the annotation of their samples with the descriptors in question. In their study descriptors were annotated by one of the authors and a random subset was independently annotated by another author (Savage et al. 2015, Supporting Information). This can hardly account for any diversification: there is no justification to the assumption that listeners with different demographic/cultural/musical background would produce similar ratings.

An example from MIR literature is a study by A. Friberg et al. (2014), in which the authors investigate a number of perceptual descriptors of music to be used as a middle layer between the low-level audio features and high-level semantic features - in a similar manner, in which we intended to use our objective descriptors of vocal production. They suggest to use perceptual descriptors such as *speed*, *rhythmic complexity*, *articulation*, *dynamics*, etc. instead of more objective categories like key or tempo. They demonstrate that these perceptual descriptors can be effective as a middle layer (correlate well to semantic descriptors). In contrast to the previous example, they rely on about 20 raters and carefully analyse the inter-rater agreement. Yet, like in the above example, and like in Cantometrics, their rating procedure does not account for the cultural diversification of raters - most of them were students at KTH in Stockholm.

Lomax also relied on this implicit assumption - that the Cantometrics ratings will only depend on the training of the raters but not on their cultural background. Our findings for vocal production do not support this assumption - as we have outlined above (Section 8.1.1.2), we found no consensus for the majority of the physiological dimensions associated with the perceptual descriptors used in Cantometrics (some of which were also the basis for the Savage et al. (2015) universals candidates).

Because these associations are hypothetical, our findings are not a rebuttal and the assumption could still hold for some Cantometrics parameters. Yet it would need further justification and research.

8.1.6 How do we revise Cantometrics

8.1.6.1 Anchoring participants' ratings

Our rich data may provide a valuable corpus to study the differences between mental representations of singing of our participants. Yet for future studies it seems more appropriate to separate this task from the task of establishing agreement about taxonomies. For the latter it would be more effective if the raters' judgements were anchored – if they were offered examples of the scale's extremes before they start their analysis. Compiling such examples in cross-cultural context is not trivial – we do not know whether we have found the most nasal vocalisation ever recorded on Earth, or the one with the widest AES. Also, the experts compiling the set of examples will be subject to their own cultural bias (see Section 7.6). This would require a collaborative effort that could be compared in scope to the Cantometrics experiment. A good first approximation would be to use the Cantometrics dataset as a widely (and to this day uniquely) representative corpus of singing from around the world and sample those extreme examples from it.

Kreiman et al. (1993) found that even experienced raters shifted along the scale during the experiment when using the type of Likert scales we used. This shift seems to reflect a real neurophysiological effect – using the visual scale which was far more fine-grained than the raters' judging capability seemed to mask this shift due to larger intervals of the scale reflecting the same rating (Kreiman et al. 1993). Providing participants with examples of scale extremes might help to eliminate this kind of shift; alternatively, other scales can be considered for future experiments.

8.1.6.2 Diversification of raters

One of our main methodological concerns about the Cantometrics experiment was participant diversification. Cantometrics raters were students of ethnomusicology from three US universities (Lomax 1977, p. 270). As we outlined in Section 2.3, while they had varying degrees of music training, Lomax fell short of introducing diversification in age, gender, occupation, experience with singing, ethnicity and cultural background. To verify universality or a wide-spread understanding of a perceptual term, such diversification is crucial.

One of the main confounding issues we discovered, discussed in Section 7.6, was the lack of familiarity with the music culture from which a sample to be analysed originated. We noticed that our experts who had experience with e.g. South East Asian musical cultures rated samples from that area differently and were able to give insights into how those sounds were made, in contrast to those raters whose main background was Western classical. Participants also sometimes explicitly stated that you have to be able to produce the sound yourself in order to teach it or to judge it. In light of this finding the question arises how well Cantometrics raters were prepared for their rating task. Given that they were mainly students of ethnomusicology it could be argued that they, among Western raters, were most experienced with sounds from other cultures and therefore most appropriate as raters. On the other hand we have no information about any of them being experienced in assessing singing which might have had a profound effect on their ratings: they would still use empathic listening (Section 7.6) as one of the instruments of their analysis, making it vulnerable to their implicit vocal habits and preventing them from taking multiple vocal strategies (Section 7.4) into account. We did not detect a professional bias though in our experiment (diversifying among singing teachers, otolaryngologists and speech pathology therapists, Section 5.6.2). So the question of diversification of raters in Cantometrics remains open. It is important to remember that Lomax did not consider diversification necessary – in his view the parameters were chosen to be as widely and easily understood as possible and his notion of a *musical core* of the culture led him to believe that main musical traits would be present in any musical utterance for everyone to detect. The good inter-rater agreement (Lomax 1977, p. 270) confirmed his argument for him. Unfortunately, this part of his thinking remains controversial and we don't see any independent evidence of this. Until his experiment is re-run with raters properly diversified we shall not know.

Diversification of raters is the main challenge in establishing a presence of consensus about a perceptual term in relation to voice production. If we are aiming at a universal or a wide consensus we have to cover not only age, gender, education and voice proficiency variation but cultural differences in the first place. To achieve that we need a large sample of participants from a large number of cultural backgrounds – ideally from around the world. An online game or a mobile app would be ideal because it would instantly be available worldwide. An attractive game element will be crucial: not only would it appeal to more people, it would also ensure that they return to the game and can be a) asked to give information

on their background, musical preferences and experience, language, age, etc. and b) presented with the same stimuli twice or more to measure intra-rater reliability. For example, a show like "The Voice" could be presented, where each player can be a judge and can choose performances that speak to him most. As a judge he can be asked to provide information on the voice using suggested terms. Examples of scientific games such as FoldIt[2] demonstrate the power of crowd-sourcing for research (Cooper et al. 2010). See Section 8.2.4.1 for further discussion.

8.1.6.3 Choosing ontology candidates

The goal of choosing right candidates for a future ontology is to ensure that there is a good chance to see listeners agree about the terms and their meanings. This thesis is an example – we chose candidates based on the knowledge of the field (Section 2) and performed the mixed-method study to verify the consensus between experts on the values of the descriptors (Section 5). We found that only two out of 11 descriptors displayed a tendency to agreement. In Section 8.1.3 we discussed the option to use both objective and perceptual descriptors in future experiments. The argument was that Lomax saw a much better agreement about his perceptual parameters than we did for our physiological ones (Section 8.1.4). We considered whether it would be reasonable to assume that some perceptual descriptors of singing might be more widely understood and better agreed upon, as Lomax believed his parameters were.

If there are in fact universal vocal descriptors which can be understood by anyone sufficiently experienced in singing, we would expect to find at least some of them in the discourse of any mature formal vocal school. Western classical singing has been formally taught and theorised for centuries and is therefore a suitable field to look for such common terms. In their excellent investigation of a linguistic corpus of French singing teachers analysing fragments of classical lyric singing Garnier et al. (2007b) collected about 600 terms describing vocal quality. About 30 of these expressions presented a high occurrence and were deemed by the authors to represent a terminology, due to the fact that many of them were known and used by all the experts of that field and were sometimes explicitly referred to as professional jargon[3].

[2] `http://fold.it/portal/index.php`, last accessed on 26/08/2017

[3] The terms given in the paper are "brillant" (bright), "clair" (light), "détimbré", "antérieur" (anterior), "bâillé" (yawned), "naturel" (natural), "soufflé" (breathy), "dans le masque" (in the mask), "ouvert" (open), "rond" (round), "nasal", "renforcement des harmoniques aigües" (reinforcement of the high harmonics), "léger" (light/weightless), "couvert" (covered), "métallique" (metallic), "soutenu" (sustained), "équilibré" (balanced), "engorgé" (engorged), "efficace" (efficient), "en appui laryngé" (with laryngeal strain), "serré" (tight), vibrato "lent" (low), "rapide"

While some of these expressions would undoubtedly refer to specific stylistic and musical features of the given musical tradition, it is among these expressions that we should be looking for candidates for universal vocal descriptors. Experienced singing teachers like P11 (who suggested her own set of descriptors) are best placed to make decisions about the choice of candidates. Without a sufficient large-scale overview of the field it would be very difficult to avoid getting stuck with endless options of more or less subjective terms and concepts (compare Section 1.2).

Cantometrics parameters could serve as candidates – Lomax found a good inter-participant agreement on them in his experiment (Section 8.1.4). Yet diversification of raters will still have to be performed to verify their wide-spread understanding (Section 8.1.6.2). Most popular tags concerning vocal production can be harvested from corpora of users annotations of music tracks.

Objective candidates can be taken from the studies like ours: our descriptors *AES*, *larynx height* (and potentially *subglottal pressure*) displayed a tendency to agreement even in our conservative setting, it is therefore to expect that they would perform at least as good in other contexts.

Another good source for ontology candidates that we might find participants agree about is vocal health literature. Perceptual evaluation of voices has been extensively used by medical professionals to assess dysfunctional voices. Expert auditory-perceptual evaluation remains the most popular mode of assessment in the clinics in spite of opportunities for objective measurements becoming widespread (Oates 2009). As opposed to our situation of a terminological chaos there are now several widely spread and agreed scales of perceptual evaluation for dysfunctional voices: GRBAS, CAPE-V and Stockholm Voice Evaluation Approach being the most widely accepted of them. Consensus about these scales contributes to their popularity though it does not guarantee their validity or reliability (Oates 2009).

A collaboration with vocal health clinicians at this stage should help to evaluate the usefulness of such perceptual descriptors for healthy voices as well as options for direct measurements of physiological descriptors.

Summarising the above we should be seeking a combination of more subjective perceptual terms and more objective ones. For the latter our successful descriptors *AES*, *larynx height* and *subglottal pressure* would be likely candidates. The best studied characteristics in vocal health are *pitch* (Houtsma 1995), *loudness* (Sundberg, Titze and Scherer 1993) which would fall into the latter category, as well as *breathiness* (Shrivastav and Sapienza 2006, Hillenbrand and Houde 1996) and

(fast), “régulier” (regular).

roughness (Dejonckere et al. 1993, Hammarberg et al. 1980, Karnell et al. 2007) that belong to the former (see Oates 2009 for further details). Garnier's (2007b) list is a further pool to look for such descriptors.

8.1.6.4 Objective descriptors – direct measurements or expert agreement

For more objective descriptors there will be various levels of sophistication of the rating procedures. For some of them such as frequency and SPL, spectral slope or noise-to-harmonic ratio ratings can be extracted directly from audio. For others direct measurement would be possible but would require the presence of the singer and more complex and expensive measurement arrangements, such as a electroglottograph recording for closed quotient or nasal stroboscopy for epilaryngeal measurements (see Kayes 2013 for a discussion of currently available measurement techniques). These arrangements would only be available for new recordings; for existing datasets like the Cantometrics dataset the approaches of the current study will have to be taken: either to achieve a high accuracy automatic classification or to establish consensus among experts on the ratings of the chosen descriptors. A combination of both approaches could be introduced, even desirable. The automatic approach, if successful, will be scalable; the challenge would be to move from new recordings for which objective measurements are available, to older ones with poor recording quality, different recording conditions and musical variability of styles now extinct. The expert consensus is difficult to achieve, as this study has demonstrated. If objective measurements of a descriptor are available, the validity of experts' ratings should be verified: whether their ratings in fact reflect the claimed physiological function. Consensus on the values does not guarantee validity; if both is confirmed on new recordings though, it can be generalised to all recordings including the ones for which no objective measurements are available (see also Section 8.2.4.2).

Because direct measurement of physiological processes nowadays usually requires expensive equipment which can only be found at clinics, pharmaceutical companies or universities, it will be medical professionals or scientists who could provide us with such measurements. If the limitations of these measurements on the process of singing are deemed acceptable for the experiment design, objective descriptors annotated with direct measurements are the most convenient data to work with.

8.1.6.5 The steps to revise Cantometrics

The steps to achieving our original aim of revising Cantometrics can now be formulated as follows:

1. choose promising c**andidates for an ontology** of vocal production from both more subjective and more objective descriptors (Section 8.1.6.3)

2. compile a set of musical **examples** representing the **extremes** of the chosen **dimensions**; put together a **training** programme for raters (Section 8.1.6.1)

3. prepare a larger **dataset of recordings** representing a variety of cultures and vocal productions to verify the universality or objectivity of the ontology terms and to collect annotations for them (in the spirit of our methodology described in Chapter 4).

4. design an experiment to verify the **intra- and inter-rater agreement** and reliability for **perceptual** descriptors (Section 8.1.6.2)

5. design an investigation of **validity** for the ratings of **objective** characteristics by experts; if this is not achievable with the current technology, **inter-rater agreement** and reliability can be measured for these descriptors in a study similar to the one presented in this thesis (Section 8.1.6.4).

6. With the **candidates** for which **consensus** could be established in steps 4 and 5, establish a **mapping** between the new ontology and the Cantometrics parameters to verify the original Cantometrics findings.

7. **Scale up**: using annotations collected in 4 and 5 for successful ontological terms design and train machine learning models to automatically rate these descriptors for any new singing recordings (Section 8.2.4.4).

The ultimate goal will be, as outlined in Section 1.4 of the Introduction, to scale up the rating procedure to include all recorded music (see 8.2.4.4). Resulting parameter vectors can then be used to measure similarity between singing samples – a measure of similarity will allow for large-scale statistical approaches to the data.

Using large-scale statistical methods we can then pursue both directions introduced by Cantometrics: geographical distribution of similar singing and correlations of our vocal production descriptors with other data, e.g. anthropological traits. The former direction could give us new insights into music migration and evolution, in

analogy with genes or languages (Grauer 2009, Grauer 2006a). The latter would allow to probe possible relationships to anthropological societal traits, similar to the original Cantometrics approach, but with many methodological issues eliminated. Because new recordings are made all the time, it will be an ongoing open-ended project, documenting and investigating not only the static map of singing, as was the case with Cantometrics, but its change over time, alongside our rapidly changing society. When we get there, one more mystery about music and singing will be lifted, be it its place in human evolution and migration or its reflection of society. Yet we are confident that our mind still has a lot for us to discover, and the new knowledge will enhance, not undermine, our ability to enjoy music and to draw inspiration from it.

8.2 For music informatics (MIR)

The original motivation for this PhD – revising Cantometrics and scaling it up to include all recordings of singing – poses a challenge deliberately exceeding current MIR capabilities. Focusing on this challenge helped us to enunciate and investigate one of the main barriers to the development of MIR: lack of annotated datasets. This thesis addresses the underlying questions behind the creation of such datasets: what are the reasons for their absence and the conditions for them to be created; when can the annotations be considered reliable and how can we collect them. We addressed the complex problem of Cantometrics revision in two ways: first, introducing a simplification and outlining a path for generalisation; second – creating new reliable annotations for the original Cantometrics dataset.

Our proof-of-concept experiment in automatic annotation of phonation modes (Chapter 3) relied on voice source characteristics from vocal acoustics as our low-level features. Because no data on phonation modes (or in fact on any aspect of vocal production in singing) was available, we created a new dataset of sustained sung vowels in all phonation modes and opened it up to MIR researchers. This enabled follow-up research on our data using a wealth of low-level features, which produced better classification results (Rouas and Ioannidis 2016, Stoller and Dixon 2016, Kadiri and Yegnanarayana 2018). Our dataset was also used in other MIR areas such as synthesis of singing. Following our approach, a new, larger dataset of isolated vocal production techniques has been published recently (Wilkins et al. 2018).

In our second study we turned to an existing corpus of singing recordings - the

Cantometrics dataset - and the question of reliability and validity of annotations. While acknowledging the achievements of Alan Lomax's project we criticised its methodology and the choice of parameters describing singing (2.3). If in the previous study the data - recordings and annotations of phonation modes - did not exist and had to be created, in this case an appropriate formal language to describe the data had to be established first. In search of more objective descriptors of vocal production we constructed an ontology of vocal production based on vocal physiology (2.2). Second, annotations had to be created and we decided for expert knowledge elicitation. We devised a methodology for collecting the experts' ratings of the ontological terms, examining their consistency and analysing problem cases (Chapter 4).

For only two out of 11 descriptors did we find good agreement among raters (Chapter 5). We concluded that, given the current state of knowledge, vocal physiology in general is not well suited as a middle layer to introduce domain knowledge into automatic classification of singing – the mechanisms to produce reliable annotations of vocal physiology are currently too limited.

For the two descriptors which displayed good inter-participant agreement we collected the reliable annotations (Table 5.7) and published the first curated cross-cultural dataset of vocal production at the Open Science Framework[4]. While it is a very small dataset, we hope that its publication will lead to follow-up research and more published data, as was the case with the Phonation Modes dataset.

We analysed the confounding issues that led to experts disagreeing about the ratings of other descriptors (Chapter 7) and discuss these and their consequences for MIR in Section 8.2.2.

Given that we did find inter-rater agreement for two of our descriptors and Lomax found a good inter-rater agreement for his perceptual parameters (Section 8.1.4) we contemplated whether a combination of objective and perceptual descriptors would be a way forward for revising Cantometrics as well as MIR approaches to singing (Section 8.1.3). Section 8.2.1 below discusses how reliable annotations can be collected for both objective and perceptual descriptors of singing.

We then turn to our original motivation – revising the Cantometrics experiment – to make it an open-ended, collaborative project supported by an online community to advance our knowledge in human evolution, population movements, the relationship between singing and other aspects of human life. The role of MIR in this endeavour is deliberated in Section 8.2.4. In particular, we found a very good

[4] `https://osf.io/pff8m/`

correlation between *larynx height* and the Cantometrics *vocal width* (5.8.1), which means that we can potentially substitute the subjective and inconsistent *vocal width* with more objective *larynx height* in the original Cantometrics setting and expect to retain the correlation to subordination of women (see Section 8.1.1.3).

This section is structured as follows: we first look in more detail at the questions related to ground truth - creation, consistency and reliability of annotations (Section 8.2.1). In this context we revisit the confounding issues that emerged from our analysis in Chapter 7, such as polysemy, cultural bias, somatic bias, visual element, and the time frame of analysis. We discuss how our findings are relevant to ground truth annotation in MIR (Section 8.2.2). We then outline the contribution of MIR to our vision of the contemporary approach to the Cantometrics experiment (Section 8.2.4). We round up with future research suggestions arising from the current work in Section 8.2.5.

8.2.1 Ground truth

A barrier to successful MIR classification algorithms for vocal production is the absence of the so-called ***ground truth*** – datasets with reliable annotations of vocal production classes or characteristics which can be used for computational model training. As our research has clearly shown, lack of annotations is a true reflection of a general **knowledge gap** about singing: its physiology, its acoustics and its perception. We know impressively little about one of the most important components of music making, the only one universally present in every human culture. We find a good anecdotal evidence of this gap in regards to MIR publications and commercial applications: while the voice is one of the main attractors for music consumption and a good singer is defining for a commercial success of a music product, only a small fraction of MIR publications are concerned with singing.

To produce reliable annotations we can either draw on **objective measurements**, **experts' annotations** or **crowd-sourcing**.

MIR researchers are familiar with **audio spectral measurements** – the most widely used kind of data in MIR which in case of singing are objective measurements of vocal acoustics. Voice science is a booming field with a lot of exciting research about vocal acoustics emerging, yet often its applications are not scalable to corpus sizes normally processed in MIR, as we found in our work. We attempted an inverse filtering approach to investigating vocal source characteristics (phonation modes, Chapter 3). Inverse filtering, a process of deconvolution of vocal source signal and

vocal tract filter from the resulting spectrum, usually requires manual adjusting of several parameters. We used an open source implementation (*TKK Aparat*) and implemented batch processing for the adjustments, yet multiple maxima were calculated for the parameters and a human judgement was still necessary. Using our own annotated dataset of sustained vowels we tested various low-level features of the voice source signal for their classification power in relation to phonation modes. While our classification attempts were statistically significant, their accuracy was not very high. Moving from sustained vowels recorded under controlled conditions to real-life recordings would have made the classification accuracy untenable.

Other **objective measurements** cover vocal **physiology** of the singers. Direct measurements can now be made in a lab with specialist hardware during the process of singing for a number of physiological characteristics. Some of measurement procedures are more invasive than others, preventing unobstructed vocalisation: e.g. measurements can only be generated for particular sound sequences like sustained vowels; or with a stroboscope camera in the singer's throat; or with the singer lying down in a noisy MRI room. We expect these technologies to improve rapidly in the coming years and become cheaper, more widespread, easier to apply and less invasive. These advances would open up ways to creating new recordings of singing with reliable annotations. It would not, however, affect our options for recordings already made.

Crowd-sourcing of music annotations is an exciting tool that has been widely exploited for MIR research. It is particularly appropriate for experiments where a large group of participants with a variety of backgrounds is needed – online questionnaires, games and apps are instantly available around the world. It seems to offer great functionality for testing agreement about perceptual/subjective descriptions of singing, as we discussed in relation to Cantometrics (Section 8.1.6.2).

Expert annotations differ favourably from crowd-sourcing in two aspects: first, a small group of experts may be sufficient as opposed to a large-scale crowd-sourcing endeavour; secondly, experts can be asked to rate technical terms which amateurs are not familiar with; in particular, they can rate physiological or acoustic descriptors which can be directly measured. These ratings would enable the researcher to measure not only consensus between raters but also the validity of that consensus: whether the ratings reflect physical reality. Even in cases where descriptors cannot be measured objectively and thus the validity of ratings cannot be established, a statistical agreement between expert listeners will be more conservative than between amateurs (experts tend to disagree more often, see Kreiman

et al. 1993) and will be a better indicator of a real common understanding. We took the expert annotation approach to physiological descriptors of singing in this thesis (see Chapter 4). We found good agreement between our participants on two descriptors out of 11 (or 17 dimensions, Chapter 5). We collected averaged ratings for these two descriptors – *AES* and *larynx height* – for the 19 snippets (and also aggregated for the 11 tracks). These annotations are reliable given the current state of knowledge on vocal production. Together with the audio they constitute the first ever cross-cultural dataset on vocal production with reliable annotations (Table 5.7).

Our extensive analysis of qualitative data (Chapter 6) allowed us to formulate confounding issues/reasons for disagreement on the other descriptors within this group of experts, which can be generalised to singing voice analysis among relevant professional groups (Chapter 7).

While currently all physiology measurement techniques are to some extent invasive and mostly very expensive, we expect, given the current pace of technological development, an emergence of a **new screening technology** or a technological leap within the existing frameworks that would bring our knowledge of vocal physiology to the next level and would allow for a more routine collection of real-life annotations of the physiology of singing. Voice acoustics will be developing in parallel offering better scalable techniques. Such changes will make MIR approaches directly applicable to the singing data. Until then automatic classification of singing will remain difficult, requiring expert knowledge from outside of MIR field.

8.2.2 Annotator's bias and how to deal with it

Commonly, in MIR, research datasets with annotations are acquired externally; often no details are provided on the process of annotation and MIR researchers have no control and sometimes little understanding about the consistency and validity of the annotations. Where no annotations are available, it seems to be common practice to self-annotate datasets: many of us are practicing musicians with a good understanding of the subject. Especially when the data in question is the voice that we all know so intimately, self-annotating seems appropriate. Yet as our work demonstrates, we have no universally understood words to describe our voices. All our voice ratings are subjective, whether we are aware of it or not, whether we are experienced in voices or not. Raters less experienced in singing voice may be tricked by seemingly self-explanatory terms like *pressed phonation* or *nasality*,

even though they can be subjective and ambiguous (see Kreiman et al. 1993); for more experienced raters, technical terms like *thyroid cartilage tilt* or *aryepiglottic sphincter narrowing* may give an impression of objectivity or precision, yet our subjective auditory-perceptual mechanisms are used to identify these characteristics.

Because our results would be as biased as our data, it is important to be aware of the bias that is inevitably introduced in the process of data annotation. We investigated in greater detail the confounding issues leading to disagreement and bias in ratings in Chapter 7. In this section we summarise our findings with the view of ground truth annotation in MIR.

8.2.2.1 Define your terms

We don't know of any widely understood terminology related to singing voice so far, therefore it has to be assumed that each and every term will be understood in a variety of ways until proven otherwise. We have pointed to the lack of common vocabulary in the Introduction (Section 1.2), did our best to define our terms in the Ontology chapter (Chapter 2), found little agreement between experts on the values of these terms (Chapter 5) and analysed polysemy and different interpretations of the terms in Section 7.1.

In our study the discrepancies arising from different interpretations of terminology are manifested in a high acceptance of our terms by the participants on the one hand and diverging ratings of their values on the other. In some cases, like *phonation* (Section 7.1), our judges were particularly clear about their ratings assigning them very high confidence, but the values did not agree, which points at them attributing different meanings to the values of the scale and thus to the term.

Sometimes terminology for a particular phenomenon was varied and confusing with various parties insisting on their version of it. A good example of this are vocal *registers*. Registration being one of the most widely used concepts, there is no commonly accepted way to talk about it. We have discussed its history in detail in Section 2.1.2 and the results of our qualitative analysis in Section 7.1. The number of registers varies from 2 to 5 depending on the source; some referring to head and chest, others to light and heavy, or simply to mechanisms M0 to M3, with falsetto meaning different things, particularly for male and female voices; For some experts registers can be mixed, for others they can not; sometimes registers are related to specific pitch ranges and in fact defined that way, other researches insist that registers are not related to pitch.

Particular care needs to be taken when subjective, perceptual descriptors are involved, because for subjective constructs no agreement can be expected a priori. To test consensus about them proper rater diversification needs to be undertaken. If consensus is only tested on a small, homogenous sample of raters, it cannot be generalised for other population groups.

8.2.2.2 Plurality of opinion

When dealing with physiological descriptors it is important to remember that our knowledge of the physical reality of vocal production is quite limited. There still are and there will be for some time different views of this reality (see Section 7.2). When you choose your taxonomy based on the view of a particular expert, you may find that other experts see the subject differently. For example, Jo Estill describes the opposition between thick and thin vocal folds (head and chest voice) as one-dimensional: vocal folds are thinner when they are elongated and thicker when they are shortened. Our participant P12 insisted that this opposition is two-dimensional: vocal folds can change between long and short and between heavy and light (Section 7.2).

In another instance P12 explained the shouty belting sound in rock music or musical theatre via thickening the vocal folds (making them heavy); for Estill belting is a result of a cricoid cartilage tilt, that would also shorten the vocal folds. Some researchers claim they have seen cricoid tilt on MRI recordings of vocal tract; others point out that there is no physiological mechanism to tilt the cricoid. These differences are significant, they reflect differing views of physical reality by the researchers. Until there is a technology allowing to screen our body's inner workings routinely in a non-invasive way during singing we will continue to see this kind of disagreements among experts.

8.2.2.3 Document raters' backgrounds

Familiarity with the kind of vocalisation one is analysing has emerged as the most important confounding issue (Section 7.6). We have seen examples of expert raters having difficulties judging unfamiliar sounds, while others who had personal experience with similar singing techniques were straightforward and confident about their ratings (Section 7.6).

The cultural context determines to a large extent what kinds of singing a person listens to and what kinds of sounds they make (Section 7.6); its influence on the

person's vocal space (Section 4.5) is paramount. In our study we found that even for very experienced voice professionals it plays a big role when they are confronted with singing traditions they are not familiar with.

Therefore collecting information on raters' cultural background and diversifying raters in this respect is crucial.

8.2.2.4 Stay with a single culture

Cultural bias is a very significant confounding issue that would affect any MIR model. If cultural differences are not subject of your experiment, you would be better advised to source your musical material from one culture/tradition. Your results would then hold for this particular tradition though, their generalisation would have to be justified or tested in further studies. We have suggested an experiment design for studying experts' agreement about vocal physiology within a single culture (Section 5.10). Performing such a study will not only be useful and informative in itself; because it is the same design as the study in this thesis, the only difference being the absence of the cross-cultural component, it will allow us to measure the effect of the cultural bias in the current study.

The ethnoMIR approach – studying non-Western musical cultures with MIR methodology – started as a small interest group about eight years ago, and has gathered momentum in recent years due to several large funded projects. In most cases each study deals with a dataset from a single culture, which is, as mentioned above, the safest approach in order not to be confounded with the cultural bias. Sometimes a comparative study is performed in which musical material from two or several cultures is present: each tradition is handled separately before they are compared. Large-scale comparative studies are very rare. This distribution is similar to ethnomusicology, though there a relativist paradigm focusing on a single culture is even more dominant and comparative approach is often viewed with suspicion. Ethnomusicologists are very aware of the cultural bias and of a possible superficiality of large comparisons learning from the rich history of their field and general humanitarian and social discourse. MIR does not have this history and our researchers, coming mainly from sciences and engineering are not well versed in humanitarian thought. Therefore, a warning sign should be on each MIR researcher's screen saver to remind us of the imminent danger of cultural bias lurking in each musical judgement we make.

8.2.2.5 Anchoring

If your study does involve multiple cultures, if your participants come from a wide range of backgrounds, or if you want to improve the agreement between them, consider anchoring their vocal spaces. This is done by means of a training session administered before the rating procedure, in which participants are presented with examples of the extremes of the scales as well as their middle (zero) positions (Section 4.5). This training also offers a good opportunity to provide verbal definitions or any other explanations about the terms.

Alan Lomax used this kind of pre-training in his Cantometrics experiment. In our mixed-methods study we decided against it, because eliciting our experts' knowledge was more important than confining them to the same vocal space (Section 4.5). For the revision of the Cantometrics experiment though we suggested to incorporate the anchoring training in order to boost the inter-rater agreement (see Section 8.1.6.1).

It is important to remember that those who are going to compile the examples for training are subject to their own cultural bias.

8.2.2.6 Reflect on the annotation process

When human listeners analyse singing, they do it via the perceptual channels available to them, in particular auditorily and visually. Yet our participants also mentioned what they called *empathic listening*, based on motor/kinaesthetic perception.

Empathy is our capacity to feel what the other person is experiencing; rapport is the ability to "sync" with the other person – in regards to singing it is when our body reconstructs the motor gestures of the other person. *Empathic listening* is evoked when the listener's body reacts to the sound of singing in that it prepares to reproduce the sound, bringing its vocal apparatus into the necessary state. The listener analyses the changes in his own body and concludes about the physiology employed by the singer (see Section 7.6). It can be a very powerful process for some individuals providing a direct access to the information about the inner bodily workings of another person. It usually happens unconsciously and particularly people who are relatively new to analysing singing are prone to relying on empathic judgements without a second thought. You need a significant experience and a strong self-reflecting ability to counteract the overwhelmingly convincing force of a strong empathy. While in some cases the judgements based on empathic listening can be accurate, in situations where the judge's physiological strategy to achieve a given acoustic result differs from the singer's errors are inevitable.

With regards to the cultural bias of the raters their judgements become particularly prone to errors when they are faced with sounds they have never heard or made. If they have no experience of making this kind of sound the body will "guess", based on its own vocal habits. An informed guess of an expert taking in account a large variety of different voices heard and analysed could plausibly be better than an unreflected guess of an amateur. Yet our study has shown that even highly experienced voice professionals fall victim to the bias if they are not in their familiar playing field culturally speaking. Self-reflection and external control is key in detecting and correcting this bias, therefore collaborative projects which have voice professionals as well as experts in the cultures with which the project engages is paramount.

8.2.2.7 Where nothing can be done

There are other aspects of singing which will introduce variance into our data which cannot be counteracted.

Difficult cases As is well-known in machine learning, in highly varied datasets (like cross-cultural singing datasets or ethnomusicological recordings) there always will be cases producing errors (see e.g. Proutskova and Casey 2009). This is not only the case for machine learning, even expert listeners struggle with some instances of singing more than with others. Many factors play into it, in particular low recording quality, environmental and other sounds overlapping with singing on the recording, absence of the visual information; all this in conjunction with a combination of physiological settings which, when they occur together, are very difficult to deconvolute and recognise auditorily. For example our participants mentioned that it is very difficult to determine whether a *falsetto* sound has *twang* (narrow *AES*, see quote in 6.7).

Another source of difficulty (and inevitable misclassification for automatic models) are vocalisations which are physiologically speaking special cases, where non-conventional physiological mechanisms are used to produce sound. We describe one such special case – the singing of Tibetan monks – that came up in the interviews, and provide a list of other singing traditions that include special case vocalisations in Section 7.3.

See Section 7.3 for further discussion of difficult cases.

Different physiological strategies In case of difficult physiological configurations or different models of physical reality the raters would normally be aware of the ambiguities in relation to their rating and other rating options. Yet there are situations where even experienced raters can be led astray. Several of our participants stressed that there are multiple physiological strategies to achieve the same acoustic outcome. For instance, in Section 7.4 we discuss our findings about all the different strategies to achieve bright sound. Our participants related brightness of the sound to narrow AES, high larynx, high tongue, spread lips (smily) or middle constrictor – one of the constrictor muscles of the pharynx.

See Section 7.4 for further discussion of different physiological strategies the singers employ to achieve the same acoustic result.

The motor/kinaesthetic perceptual channel, which our experts referred to as "empathic listening" or "projection" is particularly prone to errors of analysis in case of different physiological strategies. We discuss it below in Subsection 8.2.2.6 of this Section, or see Section 7.6 for more details.

These ambiguities – the possibility of multiple physiological realisations of a given vocal sound – will inevitably be reflected in the automatic classification models: if humans cannot classify them, the machines won't be able either.

8.2.3 Visual element

Another very important factor affecting our study was the absence of the visual element – our experts only worked with audio recordings (see Section 4.4.1). A direct implication for the study was that several physiological descriptors had to be omitted that are assessed visually (Section 4.4): the ones related to posture (body, head, neck alignment) and articulation (jaw, lips). Interestingly, our participants only occasionally mentioned the missing of the visual information. It seems that they are used to the task of purely auditory voice assessment, and they accepted the absence of the visual as part of the experiment. Another prominent issue that was nearly never mentioned was the sound quality: the ethnomusicological recordings, some of them made 50 year ago and earlier, were certainly not of the best quality. Only one participant found it to be a serious issue which in the end lead to the termination of the interview; otherwise it was barely mentioned.

It would be tremendously valuable to investigate the effect of the visual element on the decoding of physiology in particular and on MIR studies in general. Very often MIR applications rely solely on the auditory channel – processing audio recordings.

There is no conclusive evidence that results obtained in auditory MIR studies can be generalised to environments that include visual information, and vice versa. This generalisation seems to be one of the main assumptions building the basis of the MIR approach. The field now in its second decade is mature enough to open up a dialogue challenging this basic assumption.

8.2.4 Cracking Cantometrics

We discussed in section 8.1.3 that a mixed approach combining objective and perceptual descriptors would be preferable given the current state of knowledge and technology. Annotations will have to be collected through crowd-sourcing for perceptual and via measurements or expert annotations for objective descriptors. We discuss the general approach in more detail in Section 8.1.6.

In Section 8.1.6.5 we list the steps that have to be taken to revise Cantometrics. Here we shall consider each step from the MIR point of view. It will fall to MIR researchers to scale the approach to millions of recordings.

We discussed the options for choosing ontology candidates in Section 8.1.6.3. The candidates should be chosen with the knowledge of the domain but under the constrain that the inter-rater consensus on them will most probably be high.

Devising a training programme for raters to anchor their mental representations of singing is outlined in Section 8.1.6.1. Going through the training should help counteract the raters' biases about singing and equal out their inner scales. This process should facilitate agreement between raters. Since MIR researchers are subject to their own biases (e.g. cultural bias) as much as any raters are, it would be helpful if any MIR researcher involved in this experiment or generally in research about singing would go through this training.

8.2.4.1 Crowd-sourcing, consensus for perceptual descriptors

The next step in our plan to revise Cantometrics is to collect ratings for the perceptual descriptors. The main challenge is raters diversification in regards to age, gender, occupation, ethnicity, musical training, singing experience and cultural background (Section 8.1.6.2). To achieve this spectrum of diversification the number of raters will have to be large and come from different countries and different walks of life. In Section 8.1.6.2 we suggest a **game-like approach**, possibly in the form similar to "The Voice" show, where users can judge singers and help them progress in the competition, while they can earn points for rating particular aspects of con-

testants' singing or providing their personal information. It would help attract the raters and collect their ratings as well as background information on them, keeping it anonymous. MIR researchers often use games to collect annotations; there are many interface designers and HCI specialists among them and they know the tastes of their audiences, they would be in the best position to develop such a game.

One has to be very clear though what can be investigated plausibly in crowdsourcing experiments and what kind of data needs to be collected. To ensure **raters diversification**, information about their age, gender, education, singing voice proficiency, musical and cultural background has to be documented as part of the game. Yet we cannot expect the players to always give truthful answers about heir background. The game will also have to incorporate the preliminary training developed in 8.1.6.1.

Kreiman (1993) discusses the statistics to be collected to ensure **consensus among raters**. Listeners are in agreement to the extent that they make exactly the same judgments about the voices rated. Ratings are reliable when the relationship of one rated voice to another is constant (i.e., when voice ratings are parallel or correlated), although the absolute rating may differ from listener to listener. Intra-rater agreement and reliability are measured for the same rater judging the same stimuli, e.g. whether his ratings will be the same if he is presented with the same track a day later. Inter-rater agreement and reliability are calculated for different raters judging the same stimuli (Kreiman et al. 1993). If a consensus on a term can be established for a wide range of listeners, we can flag it as a widely understood or even a **universal term** (Section 8.1.3). Yet it is important to be aware of the **bias amateur** listeners tend to display: they will be able to give refined, confident judgements about voices that fall in the range they have been routinely exposed to (like speakers and singers in their culture), while all other voices they will judge as extreme (Kreiman et al. 1993).

Also, the confounding issues listed in this thesis (Chapter 7, see Section 8.2.2 of this Chapter for a summary) will be the source of bias for the raters. Because perceptual terms are usually borrowed from everyday language more people, particularly non-specialist, feel at ease with them compared to more technical objective terms; yet this acceptance often masks polysemy, interpretation and connotation ambiguities (Kreiman et al. 1993). The preliminary training discussed in 8.1.6.1 should help counteract that bias.

8.2.4.2 Objective descriptors

For physiological descriptors such as *larynx height* direct measurements can be produced in the lab during vocalisation. Because these measurements are not available for existing recordings, experts' ratings will still have to be used. In such cases the validity of experts' ratings should be examined. If experts agree about the values of the descriptor and their auditory-perceptual ratings tend to correspond well to physical reality, it can be assumed that their ratings of other recordings, for which no measurements are provided, will also reflect the reality. These ratings will be necessary for scaling up the automated rating procedure to datasets of previously unseen (new or old) singing recordings.

8.2.4.3 Temporal frame

Choosing the right time scale and frame size for analysis is an important decision for the whole experiment design: there is always a trade-off between granularity of analysis and the amount of frames to analyse (see Section 4.4.2).

In our experiments we had to negotiate a number of time scales. We began with Cantometrics tracks which are between 5 and 100 seconds long, averaging at around 30 seconds (they are fragments of longer tracks chosen by Alan Lomax for the analysis and rating of Cantometrics parameters). Because physiological setting would sometimes change significantly within such a fragment, we noticed that this time frame is too long for physiological analysis. We extracted snippets from these recording fragments in which physiology remained relatively stable. These were mainly from 4 to 13 seconds long. The snippets were the main entity of analysis in our mixed-method study: they were long enough to allow the raters to judge vocal production characteristics, but short enough to be physiologically stable (Section 4.3). Because we chose them manually their number was small (one or two per track).

There is evidence from the interviews that our experts often sought out sustained vowels and based their analysis on them (Section 7.5). A vowel change affects the tongue, the jaw and the lips; yet vocal apparatus is a highly complex structure where a small change in one element may affect many others (see quote in 6.5.2).

To account for our participants' analytical strategy that was often based on picking out sustained vowels in singing, the time frame of analysis should ideally be adjusted; physiological analysis, automatic or manual, should be preceded by an extraction of sustained vowels. Extracting sustain vowels has been done in MIR

in various contexts including cross-cultural datasets (Markaki, Holzapfel and Stylianou 2008), therefore this approach may lend itself to automatisation substituting our manual pre-processing in which we picked regions of general physiological stability. Analysing each sustained vowel may increase the number of frames by an order and would therefore only be appropriate in an automated context; prioritising some vowels over others for analysis would be a way forward for approaches involving human listeners, though automating such prioritisation would be a new challenge.

The time frame of a sustained vowel (usually less than a second) is too short for detection of perceptual qualities such as tension or nasality; it is in fact too short for any analysis based on auditory perception (including physiological analysis), therefore our participants, even if they chose to analyse a vowel, would include the context around it into their analysis. It seems that for physiological analysis a combined approach encompassing regions of overall physiological stability as well as vowel-based analysis would correspond most closely to experts' strategies. It could become quite complex though, depending on how often sustained vowels occur and how much physiology changes overall, and it would require manual pre-processing.

8.2.4.4 Scaling up

When ontology terms have been verified, the rater's consensus established and the annotations collected, the task for MIR researchers would be to train computational models to automatically rate the ontology descriptors for new singing recordings. Like any large-scale automatic classification task, it is full of risks, coming from the variance in the new data that was not present in the original training set. Examples are new recording hardware, new environmental sounds, new musical styles or languages/dialects. This challenge is even more serious for older recordings: there we encounter poor recording quality or deteriorating media, obsolete recording devices, extinct musical traditions, etc. (Proutskova and Casey 2009). These are fundamental issues of machine learning with which we are faced each time when we aim to scale up an MIR application. If this can be done the revised Cantometrics experiment will embrace all singing recordings ever made. It will be an open-ended project that MIR can really be proud of, deepening our scientific knowledge about areas like evolution, cognition and culture based on musical data (see Section 8.1.6.5 for more details).

8.2.5 Future research

Our dataset of sustained vowels (see Section 3.2) is already being used by MIR researchers for a variety of tasks, in particular for vocal synthesis. This fact highlights yet again our argument that datasets with reliable annotations are the real facilitators of new research in MIR. The dataset is publicly available under `https://osf.io/pa3ha/` and has generated some discussion. It would benefit from an independent evaluation of annotations and, most importantly, from new recordings, in particular by other singers.

We mentioned the absence of the visual aspect in our study as a significant factor (Section 8.2.3). The vast majority of MIR research is performed on audio recordings in absence of a visual component. It is based on the assumption that a visual component does not play a major role and the findings of MIR research can be generalised to any visual/non-visual context. This assumption needs to be challenged and the role of the visual aspect better understood. The same holds for recording quality.

This thesis highlighted that the current semantic chaos around vocal production is a true reflection of the knowledge gap about singing, which is hindering researchers from creating reliable annotations of singing recordings. In the absence of widely available non-invasive body screening technology allowing to produce physiological annotations during the process of singing we suggested that MIR researchers hold back waiting for such a technology to emerge. It would then produce the amount of data (annotations) on singing necessary for successful MIR investigations of large datasets of singing recordings, including deep learning approaches.

While singing artist identification has been tackled by MIR researchers, singing voice recommendation remains the million dollar question – why do listeners like particular voices and can an algorithm learn what kind of voices these are. Integrating knowledge about vocal physiology could contribute to this task. It could also facilitate genre classification based on vocal production patterns common for a given genre.

If our plans for the revision of the Cantometrics project can be realised to the stage where enough reliable annotations are collected (steps 4 and 5 in Section 8.1.6.5), MIR will be presented with a grand challenge of developing automatic classification models for scientifically verified descriptors of singing and scaling up these models to all recordings of singing: old, new and those to be produced in future (see Section 8.2.4.4). In the spirit of Cantometrics, geographic distribution

of vocal traits can then be investigated in space and time, contributing to the study of music evolution and peoples migration (see Grauer 2009). On the other side of the Cantometrics approach is detecting correlations between musical descriptors and other data e.g. anthropological societal traits, thus learning about society through learning about its music. While currently MIR mainly borrows from other fields to enrich its methodology for investigating music, such a project will put MIR in the position to use musical recordings and MIR techniques to facilitate research in other fields.

For further future research suggestions see Section 8.3.3.

8.3 For teaching singing

This thesis started firmly on the ground of music informatics and ethnomusicology, but gradually moved to the territory of interest for voice professionals, in particular singing teachers. Our research, originally motivated by a comparative approach to singing across cultures, had as its ultimate goal to teach computers to listen to vocal production from various genres and traditional cultures and to be able to compare and classify it. This goal seems to be too ambitious for the current state of technology, yet what we learnt from this research concerns vocal pedagogy directly.

We investigated the validity of a physiological approach to objectivising the language describing vocal production to the degree of formality necessary for computational design. Our primary methodology – quantitative and qualitative knowledge elicitation – relied on experts' knowledge and their ability to map physical reality.

We interviewed 13 professionals – singing teachers, ENTs, SLTs, voice scientists – with 10 to 45 years of work experience. They were presented with excerpts of singing recordings from various cultures and were asked to perform physiological analysis of the singing (Chapter 4). It was suggested to use the ontology we compiled, based mainly on Johan Sundberg's work on vocal source and on Jo Estill's physiological building blocks model (Chapter 2). Participants were not obliged to adhere to our ontology and there were lots of opportunities to give their opinion on the suggested terminology. We found that the majority of participants were familiar with most of the terms and comfortable to use them – over 80% of them rated 80% or more of our terms (Section 5.1). Yet finding consensus on the values of the descriptors turned out to be much more difficult: for only two out of eleven descriptors a tendency to agreement could be established (Chapter 5). While experts agreed about *larynx height*, *AES* and with less confidence about *subglottal pressure*, for other descriptors

such as *transglottal airflow*, *register/VF vibration mode*, *thyroid/cricoid tilt*, *velum*, *tongue position* their ratings displayed no tendency, no agreement. We have no information about the physiological reality of vocal production for recordings in our dataset. Where agreement of several professionals about the values of a physiological descriptor can be established we have a reason to assume that the agreed value reflects that physical reality (though it is not necessarily true, it is our best possible estimation); where no consensus is found, there is a clear contradiction – some of the experts did not reflect the reality of the given vocalisation correctly, and with the current state of knowledge there is no way to tell who (if anyone) got it right.

That disagreement about some of the basic physiological descriptors is prevalent among highly experienced professionals is a warning sign for the whole field of vocal pedagogy, in particular for teaching singing technique. Nowadays physiological knowledge is a must for a singing teacher and with arrival of Jo Estill's system of physiological approach to analysing and teaching singing has rightly become part of mainstream discourse in singing education. Whether singing technique is taught directly addressing physiology or assessing physiological settings is used as an intermediate step by the pedagogue to evaluate other high-level performance characteristics, our research puts the ability of the pedagogue, even a very experienced one, to deduce physiological processes taking place in the student's body, in question.

It is hard to imagine a more subjective area of teaching, which is prone to more misunderstandings than singing. We teach students to use their instrument – vocal folds, vocal tract, respiratory muscles – which neither the teacher nor the student has ever seen in its entirety, hidden within the body; the instrument so flexible and nuanced that in our technological age we are still struggling to reproduce it; the teacher uses his own instrument for demonstration and judgement, which neither he nor his student are able to see and which may differ greatly from the student's. Like in art, the teaching is shaped by the teacher's aesthetic preferences, but these are harder to pinpoint, because there is nothing visual to show and no vocabulary to explain; like in sports, it is about conditioning the muscles and optimising effort, but the result is not quantifiable. It is the auditory perception that both the teacher and the student use, but it is subjective and can differ significantly between them. There is also the matter of kinaesthetic perceptions of the student's singing which the teacher interprets by means of empathic mirroring (which is based solely on the teacher's experience and not the student's) and vice versa.

Modern technology has provided us with some valuable tools that can assist in evaluating singing, starting with a spectral view of an audio signal for a detailed

analysis of formants and timbre; a spectrogram for a real-time acoustical analysis; an electroglottograph for observing vocal folds closure cycle patterns. Yet we are still far away from being able to watch physiological structures moving in real time in an unobstructed, natural process of singing. We only have a limited access to monitoring some of these structures under laboratory conditions, often with singers' vocal tracts and/or bodies being severely impacted by the measuring hardware (see Kayes 2013 for a discussion of currently available measurement techniques). We are still not in a position to measure all the descriptors in our ontology during the process of singing. The lack of commonly understood or accepted vocabulary on vocal production is a reflection of our partial and fuzzy state of knowledge about the subject. This situation might change if a new imaging technology would allow us an unobtrusive look into the inner workings of the vocal apparatus during singing.

In the following Section 8.3.1 we shall summarise the meta-analysis (from Chapter 7) performed on the basis of our quantitative (Chapter 5) and qualitative (Chapter 6) findings as it is of direct concern for vocal pedagogy. We have focused on five confounding issues which we see as most significant reasons for disagreement between experts on the physiology of vocal production. We then proceed to the discussion of the teacher-student interaction (Section 8.3.2.1) and the future of teaching vocal technique (8.3.2.2) in the light of these findings. We wrap up with future research suggestions (Section 8.3.3).

8.3.1 Reasons for disagreement

8.3.1.1 Terminology

The terms we suggested for our ontology were well received by our participants (see Section 5.1), yet the agreement about the values of the descriptors' dimensions was low (Sections 5.5 and 5.7). The lack of agreement is a clear indicator that while the terms are understood there are discrepancies in the meanings to which these terms refer for each participant. These discrepancies may arise when terms are polysemic, often referring to related but distinct aspects of a phenomenon, sometimes with differing connotations. There are good examples in the literature on singing voice of exactly such polysemic discrepancies. Garnier et al (2007b) gives an excellent overview of several semantic uses of the term *nasality* among French classical singing teachers. Semantic analysis of the term *open throat* is given in Mitchell et al. 2003.

We aimed to choose more objective, quantifiable terms for our ontology to avoid the issue of polysemy. Yet our quantitative results show that discrepancies arising

from differing interpretations of terminology are still imminent. For instance, we expected the terms *breathy* and *pressed* phonation to be unambiguous yet our participants saw them differently. They sometimes referred to vocal health related issues, for some they were associated with inefficient, unprofessional or "incorrect" singing; and for others phonation was a non-judgemental characteristic of vocal production based on vocal source physics (Section 7.1). Cultural frame of reference was also mentioned, when singing would be judged pressed by Western standards based on aesthetic preferences.

In our ontology we introduced two descriptors related to the mode of vocal fold vibration. One reflected Estill's presentation of it, (vocal folds changing from thick to thin in the modal mode). The other one leaned on the traditional registration theory. Both descriptors were well accepted and rated by every participant, but participants' interpretations of the two descriptors varied considerably. For some they were essentially the same (thick folds = chest register). Others referred to vocal folds vibration mode as physiological and register as perceptual. Yet others saw the *thick to thin* scale as a one-dimensional description while imagining the register descriptor as a more nuanced stroboscopic picture (Section 7.1).

See Section 7.1 for further discussion.

8.3.1.2 Physiology: differing views on reality

In the previous subsection we pointed out the discrepancy between participants in the meanings they assigned to the terms in our ontology. There we related them to the polysemic use of the terms. Now we turn to other underlying reasons for the lack of consensus. One of them lies in the differing views about physical reality of vocal production. Can we or can we not tilt the cricoid cartilage? Is subglottal pressure inverse proportional to transglottal airflow? Is there a plane shift when a singer changes to falsetto? If experts disagree about these facts, their disagreement will be reflected in their ratings.

We found examples of such principal disagreements about the nature or function of physiological mechanisms in our interviews (Section 7.2). One example was the classical Estillian descriptor *vocal folds vibration mode*. In the Estill model the modal mode is one-dimensional: changing between thick and thin body of vocal folds. Yet two of our participants referred to a second dimension in the modal vibration mode, which had to do with the stiffness of the folds or the weight/mass of the vibrating part of it. Controlling the stiffness of the folds can e.g. help

counteract thinning them out when raising pitch (see Section 7.2).

Another example of disagreement about physiological mechanisms surrounded the *AES* descriptor. In the Estill model narrowing *AES* is a building block of several vocal qualities, among them *Belt*, as is common in musical theatre and CCM styles. One of our participants was adamant that belt is a vocal source phenomenon and is not produced by the filter (of which AES is part). He also mentioned middle constrictor and a possibility of non-linear aerodynamic effects resulting from its constriction. Other participants have been less radical, but some were very well aware that the epilarynx can be narrowed by means of different mechanisms, not just AES (Section 7.2, see also Section 8.3.1.4 on different physiological strategies).

8.3.1.3 Physiology: difficult cases

Some examples of vocalisation are just difficult when it comes to deconvolution of the effects that various physiological components cause (Section 7.3), and we introduced confidence ratings in our experiment to reflect that. For example, when non-linear aerodynamic effects are present, where the form (e.g. epilaryngeal narrowing) of the vocal tract filter affects the source, these are very hard to measure. These effects are also hard to feel kinaesthetically, as our participants point out – the singer feels like she is in chest while singing in low head, while thicker folds result from the back pressure of the twang (Section 7.3).

We also had a case of a participant intuitively misjudging pitch by as much as an octave due to a very high larynx position (Section 7.3). It is difficult to determine the presence or absence of AES narrowing in falsetto or the velum setting in an unfamiliar language. There are also vocalisation examples in a cross-cultural context which require separate investigations such as Tibetan monks, producing sounds so low that they cannot be generated by human vocal folds (Section 7.3).

8.3.1.4 Different strategies

In the previous two subsections we summarised reasons for disagreement between raters – differing views on physiological reality and difficult cases – of which the raters are well aware. We are now moving to the territory where this kind of awareness requires a considerable self-reflection effort. The issue in this subsection is that there are different physiological strategies to achieve the same acoustic result (see Section 7.4). A good example is the discussion around the production of bright sound.

One of our experts conducted a study on the variation in bright and dark vowels production. She stated that AES was active in most of the seven singers when producing bright vowels, but not in everybody, and that generally a large variety of strategies were used. She pointed out bringing up the larynx, lowering the velum as ways to shrink the pharyngeal space. She also called the bright vowels "smiley" referring to the lip form. Another participant, when discussing brightness in the Taiwanese example, stated that it could be the result of a high larynx, a narrow AES and/or a high tongue. Middle constrictor was also mentioned in relation to brightness. See Section 7.4 for more details.

There are also multiple ways to achieve particular physiological settings which then produce a desired acoustic outcome. For example, in the Estill system the primary way to thin the vocal folds is tilting the thyroid cartilage; yet our participants pointed out that there are other ways to do it (Section 7.4). The same is true for thickening the folds: you can adjust the crico-thyroid visor; you can stiffen the folds increasing the vibrating mass; or you can use non-linear aerodynamic effects for a thicker sound.

Our main instrument of analysis when it comes to vocal physiology is our ear. Visual evaluation also takes place but it is limited to the visible body parts, while most of our vocal apparatus is hidden. Auditory information is analysed based on our auditory experience. Additionally, it is often translated into kinaesthetic perceptions when our body recreates the process of sound production that takes place within the singer's body. We can consciously evaluate and even enhance these perceptions. Our experts call it *empathic listening* (Section 7.6). It is a very powerful instrument giving us access to what might be happening within another person's body. Yet, empathic listening, which is based on one's own bodily experiences and not the singer's, can go wrong. It is biased by one's own anatomy and experience. If you habitually use a different strategy to produce a given acoustic result than the singer, your body would reproduce your strategy, not the singer's.

Unfortunately teachers sometimes rely too readily in their analysis on what they think is their experience, but what in fact is their motor reaction to the student's singing; they intuitively expect the student to do what they themselves do and measure the student's progress based on their own production. Reading about research on different strategies or seeing it for yourself in a clinical setting is the best way to visualise the wide range of strategies and to make oneself more sensitive to possible caveats in our analysis.

8.3.1.5 Familiarity with the tradition

Another important confounding issue we faced in our study was the fact that our experts were not familiar with the musical traditions from which the singing originated. Some of them were very well aware of their bias towards Western musical culture; the subject of familiarity, of the ability to recreate the sounds came up frequently in the interviews. Below is a summary of our discussion in Section 7.6.

From the Western point of view Rating phonation, e.g. the degree of pressedness of the voice, highlighted the issue of our participants' own notion of pressedness being biased by their Western background (Section 7.6). They often pointed out that for our Western ears the sound could be characterised as pressed while in the tradition of origin that production would probably fall within the cultural norm. The difference in aesthetic preferences was also mentioned as a possible reason for discrepancies.

From personal experience Cultural bias was not limited to a Western background. P08 who teaches in an Eastern European vocal tradition was tricked by a gamelan vocalisation from Java – the singing sounded extremely high due to a very high larynx. In contrast, P01 noticed the high larynx and its effect straight away, because she had been involved with another East Asian vocal tradition – the Dhrupad from India – where the production is somewhat similar.

It was stressed that when a teacher cannot easily reproduce the sound themselves they might label it as unpleasant and wrong and that would impact their analysis. It is often the case at Western musical higher education institutions that singing teachers with a classical background have to teach musical theatre or CCM, which they cannot sing, and this is very unfortunate, our participants insisted. Personal experience with particular sounds and techniques emerged as a crucial advantage with regards to physiological analysis in our interviews (Section 7.6).

Language and phonetics Language and phonetics were an important factor that was often mentioned (7.5). Singers shape the vowels in the same way as they do in speech and that has a direct and crucial impact on their singing. Without any familiarity with the phonetics of the singer's speech it is very hard to rate the degree of deviation of his physiological setting from the norm. In particular, an appropriate level of subglottal pressure, a neutral larynx height and velum position were suggested to be determined by the language phonetics.

Tradition vs physiology The musical tradition in which we grow up and live shapes us as singers and as listeners. We acquire our aesthetic preferences as well as performance practice standards from our culture. Vocal tradition serves as a three-fold filter: a) it sets the standards of vocal production for the singer and the audience, b) the sounds that are culturally preferred are produced more often, therefore physiological settings necessary for their production are employed regularly and are trained, c) singers that can produce culturally preferred vocalisation more easily or skilfully are favoured against all other voices (Section 7.6). To some extent a singing tradition becomes self-reproducing.

Empathic listening Empathic listening is our ability as listeners to "feel what the singer feels" – to mirror the singer's motor gestures onto our body and vocal apparatus. We mentioned above how it can lead to mis-interpreting physiology in relation to different physiological strategies leading to the same acoustic outcome (Section 8.3.1.4). In case of analysing the sounds/traditions the listener is not familiar with, the pitfalls related to the motor mirroring are as important: their body will "guess" how to produce these unfamiliar sounds, based on the listener's vocal habits and experience. We discuss empathic listening in more detail in Section 7.6.

8.3.2 Discussion

After discussing possible reasons for disagreement between experts about the physiology they hear, we shall analyse how they affect the teacher/student interaction and manifest themselves in the studio. We shall consider our current limitations in teaching vocal technique and how a new kind of imaging technology could change our teaching.

8.3.2.1 Teacher and student

The above issues that crystallised from our study give us an insight into why experienced listeners may disagree about physiological analysis. Now we would like to examine what their impact would be on the teacher-student communication. As teachers we constantly analyse the student's physiology, sometimes for a direct correction, but more often as a means of a further high-level analysis. As we have shown in this thesis, even a very experienced teacher can go wrong with their analysis and for most traits there is currently no way to independently corroborate

our conclusions on physiology in the studio. We are biased by our anatomy, our experience, our aesthetic preferences and our culture. Students more often pick up terminology from their teachers and in a longer teacher-student relationship there is enough time to arrive at shared meanings of the teacher's terms; it is still important to remember that most terms a singing teacher uses are subjective and may have a very different meaning for the student (Section 8.3.1.1). Imagery is a valuable instrument of teaching and often more powerful than referring directly to physiology to which the student has no direct access; it is crucial that the teacher understands the difference between precise vocabulary and imagery and when the former needs to be used. It has to be mentioned that even using objective and unambiguous terms for our instructions does not guarantee understanding, because the student's control of his physiology is limited, particularly at the early stages of his education, and we as teachers again rely on our auditory-perceptual analysis for our feedback which is biased and can be inaccurate.

When we do give direct physiological instructions it is important that we have done our best to convince ourselves in the validity of our views on physiology. If we ask a student to tilt their cricoid cartilage, we may be asking them to perform an impossible act of bodily movement. Given our incomplete knowledge of vocal production mechanisms there is always a danger that what we hold true today will be shown to be false tomorrow (Section 8.3.1.2). As teachers we are obliged to follow new developments in science with regards to vocal physiology. Yet new findings will not always be independently confirmed and contradicting evidence is common in science. We have to use our judgement when incorporating new knowledge in our teaching; but we cannot afford to base our methods on anecdotal evidence or our own experience only.

The different strategies argument (Section 8.3.1.4) adds another level of uncertainty to the anatomical differences between the teacher and the student. While abdominal muscles or head position can be visible, if inner structures such as AES, velum or ventricular folds are involved we may not even notice that the student's strategy to produce a particular acoustic outcome differs from ours. We are only equipped with our ears and our kinaesthetic reactions, the latter based on our own bodily experiences not the students (Section 7.6).

Commonly the student and the teacher share a cultural background and the musical tradition/genre that is taught. Yet if our career involves crossing the boundaries of musical traditions we experience an added level of bias (Section 8.3.1.5). Our aesthetic preferences are shaped by our culture and it is impossible to avoid

imposing our judgements about the sound or interpretation on our students. As we have shown in this thesis, our cultural bias can have a significant effect on the accuracy of physiological analysis. Personal experience with the sounds we teach and analyse seems to play an important role, so a good rule of thumb seems to be "only teach what you can sing".

Teaching singing is much more than just correcting and developing vocal technique. All teachers have their strengths. Understanding one's limitations and what is best for each student is key.

8.3.2.2 Teaching vocal technique

This thesis demonstrated that even very experienced voice professionals hardly agree about physiological processes in singing (on the example of our cross-cultural dataset) and discussed possible reasons for the disagreement. We cannot objectively measure the values of the physiological descriptors for the recordings in our dataset, therefore the validity of the ratings cannot be established. Yet since there is no consensus on all but two dimensions, some of the experts will have scored closer to physical reality than others. That means some of them got it wrong. This is an uncomfortable notion, since all our interviewees have decades of experience in their profession, analysing voices routinely many times a day. Many of them are in senior positions at their institutions and/or are acknowledged internationally as leading figures in their fields. If these people can not get it right, who can?

Our pool of experts represents three occupations: otolaryngologists/surgeons, speech and language pathologists, singing teachers (a good number of participants were also voice scientists). We didn't find a better consensus on physiology within professional groups (Figure 5.6.2). We also had two groups of influence within the pool: several experts were close in their research interests and methods to Johan Sundberg; another subgroup were professionals informed by or influenced at some point in their career by Jo Estill's approach. Again, limiting the experts pool to one of the influence groups did not increase the agreement among their ratings (Figure 5.6.2).

In their paper analysing the vocabulary of classical singing teachers in France Garnier et al. note that teachers use sound and physiological analysis as means to progress to a higher level analysis of expression and phrasing (Garnier et al. 2007b). But what if their analysis is not correct? Does it really matter if their opinions about the student's physiological processes does not reflect the reality?

Our work has shown that this is probably the case more often than we would like to believe and that this can happen even to very experienced teachers. If we all can be wrong about physiology, it is important that we are aware of our limitations in analysing physiology. We are biased by our anatomy and experience which we cannot impose on our students. Their physiological strategies may be different from ours. The more anatomical and cultural differences there are between the teacher and the student the higher the probability of a misjudgement.

This situation puts our profession in a dubious light and questions our ability to teach vocal technique. There is an argument that if anyone can be wrong and no one knows the right answer, we don't need to care too much about being correct in our analysis as long as we get good vocal results with our students in spite of the analytical weaknesses we might have. While singing education is much more than just vocal technique and getting good results for the students is priority (whatever "good results" might mean for each of us or our students), in terms of teaching vocal technique we are now in a position similar to medicine of about 150 years ago, before X-rays were invented. We do not know how much better we would be able to help our students (and ourselves) if we had instruments to screen our body during singing in a non-invasive way or otherwise measure physiological descriptors of interest objectively. Given the pace of technological development, such screening technologies might be available to us in a near future. Having such a technology at hand would revolutionise our knowledge of vocal production and change our teaching practice thoroughly. We can only brace for the change and hope that exciting times that lie ahead will come within our lifetime.

8.3.3 Where to go from here – future research

During the course of this PhD we have collected a wealth of data and experience, which invite to continuation of this line of research. The first step would be to evaluate the data we have already collected, for example to analyse confidence values our participants provided with their ratings – to find out which descriptors or which tracks have caused more uncertainty and why. Salience numbers can also be studied, providing an insight into which descriptors were more important or easier to rate, and which could be inferred. They might even shed light on what was special about the three descriptors – *larynx height*, *AES* and *subglottal pressure* – that enabled the experts to arrive at a consensus about those in contrast to all other descriptors.

At the end of Chapter 5 we suggest an experiment that would help estimate the

extent of cultural bias that affected the experts during the current study. While in this study we deliberately chose musical examples from a large number of cultures making sure that our participants are generally not familiar with them, in the suggested experiment the goal is the opposite: to make sure that all participants are equally well versed in the genres/traditions represented by musical examples (Section 5.10). A comparison of quantitative agreement results from the current study and the suggested experiment would shed light on the extent of cultural bias on agreement about physiology.

For our study design we chose snippets – vocal fragments which are largely physiologically stable of about 2 seconds or more in duration – to be the main entities of analysis. The main reason for this was that this time scale was necessary to collect other perceptual information we needed. We have discussed the granularity of analysis in detail in Section 4.4.2. We found evidence in our interviews that experts, singing teachers in particular, analysed different vowel sounds separately, even within a snippet. For future experiments the question of choosing the most suitable time scale and entity of analysis should be considered carefully (see Section 8.2.4.3).

The area that came increasingly to our attention during our study and that requires further research is what one of our participants called *empathic listening* – the rapport phenomenon when the listener's body aligns with the singer's and reconstructs on the motor level the physiological process of sound production the singer used. It seems to be an important mechanism in analysing physiology of singing. We have discussed advantages and possible caveats of this mechanism in Section 7.6. The debate about motor theory of speech perception has been ongoing for decades, but we have not found any relevant research on this phenomenon in relation to physiological analysis of singing, and the one paper on auditory-vocal mirroring (Prather et al. 2008) does not cover analysis of vocal sounds. The experiment could involve singing under lab conditions with e.g. a stroboscope camera inserted through the nose of the singer; and experienced singing teachers analysing the physiology of the sound as well as reflecting about their analytical process. Their analysis can then be compared with stroboscopic pictures to evaluate the validity of their analysis.

Further discussion of future research directions can be found in Section 8.2.5.

Revising Cantometrics Singing teachers' and clinical voice professionals' expertise would be indispensable in order to reach our original goal – to revise the Cantomet-

rics experiment. In Section 8.1.6.5 we list the steps necessary to reach that goal. It emerged from our discussion in 8.1.3 that to solve the case in general we might rely on a combination of perceptual and objective descriptors. The first step – the choice of suitable candidates for the ontology – should be performed by vocal production professionals. They can draw terms that are likely to display a good raters' consensus from our study, where three physiological descriptors (*AES*, *larynx height*, *subglottal pressure*) performed well. Our participants also suggested other, more subjective sets of terms (see Section 8.1.3). Widely understood vocabulary might also be borrowed from existing terminology of well-established vocal traditions with formal training, such as Western classical. See Section 8.1.6.3 for further discussion.

The second step – devising training for raters to anchor their mental representations of singing – requires a collaboration of singing experts from different cultures and traditions, including ethnomusicological archives curators and traditional music collectors. The most challenging part of this cross-cultural endeavour will be to choose examples of singing representing the extremes of the ontological dimensions from the previous step (see Section 8.1.6.1). Voice professionals contribution will also be required for the third step – compiling a larger dataset of singing examples representing different cultures and a variety of vocal productions. Preprocessing of these examples may be necessary in the spirit of Section 4.3.

The fifth step will involve clinicians to perform measurements of objective vocal descriptors. For those descriptors which cannot yet be measured in a non-invasive way a consensus between experts will be sought through an approach similar to this study, involving interviews with experts (including singing teachers) and their mixed-method evaluation (Section 8.1.6.4).

Most importantly, if statistical results can be achieved by machine-learning experts in a large-scale experiment involving singing (see Section 8.1.6.5), their interpretation in real life, their justification or refutal, and follow-up research will be the prerogative of ethnomusicologists, voice scientists, music psychologists and voice professionals.

Appendix

All the data that was generated and analysed for this thesis is stored at the Open Science Framework repository for long-term preservation. It allows for controlled access and collaboration on the data as well as preservation. All our data is publicly available with the exception of the audio recordings for which we have no permission as well as personal information about the study participants.

Two projects were created: Phonation Modes Dataset and Vocal production ontology.

Phonation Modes Dataset

`https://osf.io/pa3ha/`

There are four datasets: two recording sessions were documented with two different hardware sets simultaneously.

One of these datasets – recorded with the N/D357A microphone – was further edited in preparation for inverse filtering: it was cut into single vowel recordings with beginnings and ends of the phonation trimmed. This prepared dataset was used in our phonation modes experiment in Chapter 3.

There is also an extensive documentation, including a Wiki page describing the dataset in detail, manuals for the hardware that was used for recordings as well as a description of hardware settings that were used.

The data is available under the Creative Commons BY-NC-SA licence.

Further Wiki pages have been set up: Discussion, Research using the dataset and Related research. Some discussion has taken place and two follow-up experiments have been presented.

A component of the project called "Automatic phonation mode classification with inverse filtering" contains our Matlab code and a detailed description on how to replicate our experiment.

Vocal production ontology

`https://osf.io/pff8m/`

This project contains all the data from our mixed-method study described in Chapters 4, 5, 6 and 7. It has a complex structure because some modules hold private data.

The component "Musical examples" contains 11 audio recordings from the Cantometrics Training Tapes dataset that we used for this research (see Figure 4.3.2). It also holds the physiologically stable snippets from these tracks that were analysed by our participants in the interviews (see Section 4.3). Annotations that are prior to our study as well as our physiological analysis of the tracks are also part of this component. Another important aspect are the licences regulating the use of the tracks.

The "Methodology" component is comprised of various documents created during the study design and planning stage (see Chapter 4). There are also the templates for the consent form, the physio form for collecting quantitative data and the interview protocol.

The "Interviews" component holds for each of the participants the audio recordings of the interview, a transcription where one was made, a background précis where available and the participant's consent form. This information is not public but can be requested from us if necessary.

"Quantitative data" contains all the ratings data collected in the interviews (see Section 4.6). It has all the original data tables for each interview/participant (which include confidence ratings); aggregated tables for each descriptor; perception ratings. We also included our own ratings of the same tracks which we did as an exercise in preparation to the study (this data was not used in the study).

"Krippendorff's bootstrapping in R" contains our R code that implements Prof. Krippendorff's bootstrapping algorithm (see Section 5.4). krippalpha.boot.R is the first implementation exactly following Prof. Krippendorff's algorithm. It is the implementation used in the kripp.boot() R package available on GitHub at `https://github.com/MikeGruz/kripp.boot`. krippalpha.boot_V2.R is the second version which extends the original algorithm to weighted ratings (Section 5.7). It introduces a new metric (alongside nominal, ordinal, interval and ratio) called "confidence": instead of a usual observations matrix the function takes a complex number matrix where the real part is a rating and the imaginary part is the corresponding confidence value. Also, independent tests that were performed to the first version are included.

"Inter-rater agreement in R" holds our implementation of the experiment. Historical code and results are included as well as a ready-to-run folder with all the code and data necessary to replicate the experiment.

"NVivo" is our NVivo project file. It was created with NVivo 10 Mac version and can be accessed upon request. It contains all our qualitative analysis including open coding.

"Vocal Production Dataset" links together the Cantometrics audio recordings, their licences and the averaged ratings for the two descriptors which displayed experts' consensus: *AES* and *larynx height*.

Bibliography

Airas, Matti (2008). "TKK Aparat: An environment for voice inverse filtering and parameterization". In: *Logopedics Phoniatrics Vocology* 33, pp. 49–64.

Alku, P. (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering." In: *Speech Commun.* 11, pp. 109–118.

Bailly, Lucie, Nathalie Henrich and Xavier Pelorson (2010). "Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling a". In: *The Journal of the Acoustical Society of America* 127.5, pp. 3212–3222.

Baken, Ronald J (2006). "An overview of laryngeal function for voice production". In: *Vocal Health and Pedagogy, Volume: Science and Assessment* 1, p. 65.

Barik, H.C. (1977). "Cross-linguistic study of temporal characteristics of different types of speech material". In: *Language and Speech* 20, pp. 116–126.

Bartholomew, W. (1934). "A physical definition of "good voice quality" in the male voice". In: *J Acoust Soc Am.* 6, pp. 25–33.

Bartmann, Manfred (1994). "Rauhigkeiten in der Volksmusik in der kanarischen Insel El Hierro". In: *Berichte aus dem ITCM-Nationalkomitee Deutschland*. Ed. by M Böckner. Vol. 3. Bamberg.

Bateman, Laura Anne (2010). "Parallels between singing and phonetic terminology". In: *Working Papers of the Linguistics Circle* 15, pp. 79–84.

Behnke, E. (1880). *The mechanism of the human voice*. 12th ed. London: J. Curwen and Sons.

Bell, Cindy L. (2004). "Update on Community Choirs and Singing in the United States". In: *International Journal of Research in Choral Singing* 2.1.

Benson, Elizabeth Ann (2017). *The Estill voice model: theory & translation.*

Birch, Peer et al. (2002). "Velum Behavior in Professional Classic Operatic Singing". In: *Journal of Voice* 16.1, pp. 61–71.

Borch, D. Zangger and Johan Sundberg (2011). "Some phonatory and resonatory characteristics of the rock, pop, soul, and Swedish dance band styles of singing". In: *J Voice* 25.5, pp. 532–7. DOI: 10.1016/j.jvoice.2010.07.014.

Bradley, Kevin, ed. (2009). *Guidelines on the Production and Preservation of Digital Audio Objects: Standards, Recommended Practices and Strategies (IASA-TC 04)*. 2nd ed. IASA (International Association for Sound- and Audiovisual Archives) Technical Committee.

Butte, Caitlin J et al. (2009). "Perturbation and nonlinear dynamic analysis of different singing styles". In: *Journal of Voice* 23.6, pp. 647–652.

Castellengo, Michèle, Bertrand Chuberre and Nathalie Henrich (2004). "Is voix mixte, the vocal technique used to smoothe the transition across the two main laryngeal mechanisms, an independent mechanism". In: *Proceedings of the International Symposium on Musical Acoustics*.

Celma, Òscar and Xavier Serra (2008). "FOAFing the music: Bridging the semantic gap in music recommendation". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.4, pp. 250–256.

Chang, Chih-Chung and Chih-Jen Lin (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapman, Janice L (2011). *Singing and teaching singing: A holistic approach to classical voice*. Plural Publishing.

Childers, DG and C-F. Wong (1994). "Measuring and modeling vocal source-tract interaction." In: *IEEE Trans Biomed Eng.* Vol. 41. 66371.

Cleveland, Thomas F, P Johan Sundberg, Jan Prokop et al. (2003). "Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer". In: *Journal of Voice* 17.3, pp. 283–297.

Colton, Raymond H and Jo A Estill (1981). "Elements of Voice Quality: Perceptual, Acoustic, and Physiologic Aspects". In: *Speech and language: advances in basic research and practice*. Ed. by Norman J. Lass. Vol. 5. New York: Academic Press, pp. 311–403.

Cooke, Nancy J (1999). "Knowledge elicitation". In: *Handbook of applied cognition*, pp. 479–510.

Cooke, Peter (2006). "Response to Echoes of Our Forgotten Ancestors". In: *World Of Music* 48.2.

Cooper, Seth et al. (2010). "Predicting protein structures with a multiplayer online game". In: *Nature* 466.7307, pp. 756–760.

Creswell, John W (2006). *Qualitative Inquiry And Research Design: Choosing Among Five Approaches Author: John W. Creswell, Publisher: Sage Publica*. Sage Publications, Inc.

Cross, Ian (2006). "Four issues in the study of Music Evolution". In: *World Of Music* 48.3.

– (2012). "Music and biocultural evolution". In: *The cultural study of music*. Routledge, pp. 39–49.

Cruttenden, A. (2014). *Gimson's Pronunciation of English*. Routledge.

Curwen, J. Spencer (2010). *The Boy's Voice. A Book of Practical Information on The Training of Boys' Voices For Church Choirs*. Originally published in 1891. The Project Gutenberg. URL: `http://www.gutenberg.org/ebooks/32023`.

Dejonckere, PH et al. (1993). "Perceptual evaluation of dysphonia: reliability and relevance". In: *Folia Phoniatrica et Logopaedica* 45.2, pp. 76–83.

Drugman, T. et al. (2008). "Glottal source estimation robustness". In: *Proc. of the IEEE International Conference on Signal Processing and Multimedia Applications (SIGMAP08)*.

Drugman, Thomas, Baris Bozkurt and Thierry Dutoit (2012). "A comparative study of glottal source estimation techniques". In: *Computer Speech and Language* 26, pp. 20–34.

Dunbar, Robin IM (2012). "On the evolutionary function of song and dance". In: *Music, language, and human evolution*, pp. 201–14.

Echternach, Matthias et al. (2008). "Vocal tract and register changes analysed by real-time MRI in male professional singers—a pilot study". In: *Logopedics Phoniatrics Vocology* 33.2, pp. 67–73.

Esling, John H, Jimmy G Harris and J Romero (2003). "An expanded taxonomy of states of the glottis". In: *Proceedings of the 15th international Congress of Phonetic Sciences*. Vol. 1, pp. 1049–1052.

Estill, J and RH Colton (1979). "The identification of some voice qualities". In: *The Journal of the Acoustical Society of America* 65.S1.

Estill, Jo (1988). "Belting and classic voice quality: some physiological differences". In: *Medical problems of performing artists* 3.1, pp. 37–43.

Estill, Jo et al. (2005a). "Estill Voice Training: Level One, Figures for Voice Control". In: *Estill Voice Training Systems International, LLC*.

– (2005b). "Estill Voice Training: Level Two, Figure Combinations for Six Voice Qualities". In: *Estill Voice Training Systems International, LLC*.

Fadiga, Luciano et al. (2002). "Speech listening specifically modulates the excitability of tongue muscles: a TMS study". In: *European Journal of Neuroscience* 15.2, pp. 399–402.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague, Netherlands: Mouton.

Feld, Steven (1984). "Sound Structure as Social Structure". In: *Ethnomusicology* 28.3, pp. 383–409. ISSN: 00141836. URL: http://www.jstor.org/stable/851232.

Födermayr, Franz (1971). *Zu gesanglichen Stimmgenbung in der außereuropäischen Musik. Ein Beitrag zur Methodik der vergleichenden Musikwissenschaft*. Wien: Stieglmayr.

Ford, David N and John D Sterman (1998). "Expert knowledge elicitation to improve formal and mental models". In: *System Dynamics Review* 14.4, pp. 309–340.

Friberg, Anders et al. (2014). "Using listener-based perceptual features as intermediate representations in music information retrieval". In: *The Journal of the Acoustical Society of America* 136.4, pp. 1951–1963.

Fritzell, Bjorn (1992). "Inverse Filtering". In: *Journal of Voice* 6.2, pp. 111–114.

Froeschels, Emil (1943). "Hygiene of the voice". In: *Arch Otolaryngol.* 38.2, pp. 122–130.

Galantucci, Bruno, Carol A Fowler and Michael T Turvey (2006). "The motor theory of speech perception reviewed". In: *Psychonomic bulletin & review* 13.3, pp. 361–377.

Garcia, Manuel (1847). *MeÂŽmoire sur la voix humaine preÂŽsenteÂŽ a' l'AcadeÂŽmie des Sciences en 1840*. 2nd ed. Paris: Imprimerie d'E. Duverger.

García, Manuel (1855). "Observations on the Human Voice". In: *Proceedings of the Royal Society of London*. Vol. 7, pp. 399–410.

García, Manuel (1884). *TraiteÂŽ complet de l'Art du Chant*. 8th ed. Paris: Heugel et Cie.

Garnier, Maëva et al. (2007a). "Characterisation of Voice Quality in Western Lyrical Singing: from Teachers' Judgements to Acoustic Descriptions". In: *journal of interdisciplinary music studies* 1.2, pp. 62–91.

Garnier, Maëva et al. (2007b). "Characterisation of Voice Quality in Western Lyrical Singing: from Teachers' Judgements to Acoustic Descriptions". In: *Journal of interdisciplinary music studies* 1.2, pp. 62–91.

Granqvist, Svante (2003). "Computer methods for voice analysis". PhD thesis. Department of Speech, Music and Hearing, Stockholm.

Granqvist, Svante et al. (2003). "Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup". In: *Journal of Voice* 17.3, pp. 319–330.

Grauer, Victor A. (2006a). "Echoes of Our Forgotten Ancestors". In: *The World Of Music* 48.2.

– (2006b). "Echoes of Our Forgotten Ancestors: Some Points of Clarification". In: *The World Of Music* 48.3.

– (2007). "New perspectives on the Kalahari debate: a tale of two 'genomes'". In: *Before Farming, the archaeology and anthropology of hunter-gatherers* 2.

– (2009). "Concept, Style, and Structure in the Music of the African Pygmies and Bushmen: A Study in Cross-Cultural Analysis". In: *Ethnomusicology*.

Gudnason, Jon, Daniel P.W. Ellis Mark R.P. Thomas and Patrick A. Naylor (2012). "Data-driven voice source waveform analysis and synthesis". In: *Speech Communication* 54, pp. 199–211.

Guzman, Marco et al. (2015). "Laryngoscopic and spectral analysis of laryngeal and pharyngeal configuration in non-classical singing styles". In: *Journal of Voice* 29.1, 130–e21.

Hammarberg, Britta et al. (1980). "Perceptual and acoustic correlates of abnormal voice qualities". In: *Acta oto-laryngologica* 90.1-6, pp. 441–451.

Harris, Thomas Martin, Sara Harris and John Stephen Rubin (1998). *The voice clinic handbook*. Whurr Publishers Limited.

Helmholtz, Hermann von (1877). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. F. Vieweg.

Henrich, Nathalie (2006). "Mirroring the voice from Garcia to the present day: some insights into singing voice registers". In: *Logopedics Phoniatrics Vocology* 31.1, pp. 3–14.

Henrich, Nathalie et al. (2006). "Period-doubling occurrences in singing: The 'bassu'case in traditional Sardinian 'A Tenore'singing". In: *ICVPB (2006), Tokyo*.

Hillenbrand, J and RA Houde (1996). "Acoustic correlates of breathy vocal quality: Hillenbrand J, Houde RA: Acoustic corre- tion of dysphonia. J Voice 1999;13:508–517. lates of breathy vocal quality: dysphonic voices and continuous speech". In: *J Speech Hear Res* 39, pp. 311–321.

Hirano, Minoru (1974). "Morphological structure of the vocal cord as a vibrator and its variations". In: *Folia Phoniatrica et Logopaedica* 26.2, pp. 89–94.

– (1981). *Clinical Examination of Voice*. New York: Springer.

– (1988). "Vocal mechanisms in singing: laryngological and phoniatric aspects". In: *Journal of Voice* 2.1, pp. 51–69.

Hirano, Minoru and Yuki Kakita (1985). "Cover-body theory of vocal fold vibration". In: *Speech science*, pp. 1–46.

Hollien, Harry (1974). "On vocal registers". In: *Journal of Phonetics* 125.2.

– (1983). "Report on vocal registers". In: *Musical Acoustic Conference (SMAC)*. Ed. by A. Askenfelt et al. Vol. 46. 1. Stockholm, Sweden: Royal Swedish Academy of Music, pp. 27–35.

Houtsma, Adrianus JM (1995). "Pitch perception". In: *Hearing* 6, p. 262.

Howard, D. M. et al. (2004). "Winsingad: A Real-Time Display for the Singing Studio". In: *Logopedics Phoniatrics Vocology* 29.3, pp. 135–144.

Howard, David M. (2010). "Electrolaryngographically revealed aspects of the voice source in singing". In: *Logopedics Phoniatrics Vocology* 35.2, pp. 81–89.

Huron, D. (2001). "Is music an evolutionary adaptation?" In: *Annals of the New York Academy of Sciences* 930, pp. 43–61.

Kadiri, Sudarsana and Bayya Yegnanarayana (2018). "Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)". In: *Proc. Interspeech 2018*, pp. 441–445.

Karnell, Michael P et al. (2007). "Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders". In: *Journal of Voice* 21.5, pp. 576–590.

Kayes, G. (2004). *Singing and the Actor*. Routledge. ISBN: 9780878301980.

Kayes, Gillyanne (2013). "How does genre shape the vocal behaviour of female singers?" PhD thesis. London: Institute of Education, University of London.

Kelso, JA Scott (1997). *Dynamic patterns: The self-organization of brain and behavior*. MIT press.

Kob, Malte et al. (2011). "Analysing and Understanding the Singing Voice: Recent Progress and Open Questions". In: *Current Bioinformatics* 6, pp. 362–374.

Kreiman, Jody et al. (1993). "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research". In: *Journal of Speech, Language, and Hearing Research* 36.1, pp. 21–40.

Krippendorff, Klaus (2012). *Content analysis: An introduction to its methodology*. Sage.

Lamesch, Sylvain et al. (2007). "Investigating voix mixte: A scientific challenge towards a renewed vocal pedagogy." In: *3rd Conference on Interdiciplinary Musicology, CIM07*.

Large, John (1972). "Towards an integrated physiologic-acoustic theory of vocal registers". In: *The NATS Bulletin* 28, pp. 30–35.

Laver, John (1980). "The phonetic description of voice quality". In: *Cambridge Studies in Linguistics London* 31, pp. 1–186.

– (2009). *The Phonetic Description of Voice Quality*. Language Arts and Disciplines. Cambridge University Press.

Laver, John David Michael Henry (1987). "Individual features in voice quality". PhD thesis. University of Edinburgh.

Lehto, Laura et al. (2007). "Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types". In: *J Voice* 21.2, pp. 138–50. DOI: 10.1016/j.jvoice.2005.10.007.

Lichte, William Heil (1941). "Attributes of complex tones." In: *Journal of Experimental Psychology* 28.6, p. 455.

Lieberman, Philip (1998). *Eve spoke: Human language and human evolution*. WW Norton & Company.

Little, M.A., Declan A.E. Costello and Meredydd L. Harries (2009). "Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures". In: *Journal of Voice*.

Lomax, Alan (1968). *Folk Song Style and Culture*. New Brunswick, New Jersey: Transaction Books.

– (1977). *Cantometrics: A Method of Musical Anthropology (audio-cassettes and handbook)*. Berkeley: University of California Media Extension Center.

Lombard, Lori E and Kimberly M Steinhauer (2007). "A novel treatment for hypophonic voice: Twang therapy". In: *Journal of Voice* 21.3, pp. 294–299.

Magas, Michela and Polina Proutskova (2013). "A location-tracking interface for ethnomusicological collections". In: *Journal of New Music Research* 42.2.

Markaki, Maria, Andre Holzapfel and Yannis Stylianou (2008). "Singing Voice Detection using Modulation Frequency Features". In: *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*.

Mayerhoff, Ross M et al. (2014). "Analysis of supraglottic activity during vocalization in healthy singers". In: *The Laryngoscope* 124.2, pp. 504–509.

McGlashan, Julian (2013). "What descriptors do singing teachers use to describe sound examples?" Presented at PEVOC 10 (Pan-European Voice Conference) Prague, Czeck Republic.

Merker, Björn (2000). "Synchronous Chorusing and Human Origins". In: *The Origins of Music*. Ed. by Nils Wallin, Björn Merker and Steven Brown. Cambridge, MA: MIT Press.

Merriam, Alan P. (1969). "Review of A. Lomax, Folk song style and culture." In: *The Journal of American Folklore* 82.326, pp. 385–387. ISSN: 00218715, 15351882. URL: http://www.jstor.org/stable/539790.

Miller, Jeffrey (2000). "Evolution of Human Music through Sexual Selection". In: *The Origins of Music*. Ed. by Nils Wallin, Björn Merker and Steven Brown. Cambridge, MA: MIT Press.

Mitchell, Helen F. et al. (2003). "Defining 'open throat' through content analysis of experts' pedagogical practices". In: *Logopedics Phoniatrics Vocology* 28.4, pp. 167–180. DOI: 10.1080/14015430310018856. eprint: http://informahealthcare.com/doi/pdf/10.1080/14015430310018856. URL: http://informahealthcare.com/doi/abs/10.1080/14015430310018856.

Mithen, Steven (2005). *The Singing Neanderthals: The Origins of Music, Language, Mind and Body*. London: Weidenfels and Nicolson.

Mithen, Steven and Nicholas Bannan (2004). *Music, Language and Human Evolution (EXPLORATORY WORKSHOP), scientific report*.

Moisik, Scott (2008). "A three-dimensional model of the larynx and the laryngeal constrictor mechanism: Visually synthesizing pharyngeal and epiglottal articulations observed in laryngoscopy". PhD thesis. university of Victoria.

Moisik, Scott and John H Esling (2011). "The 'whole larynx'approach to laryngeal features". In: *Proceedings of ICPhS, Hong Kong*, pp. 1406–1409.

Moisik, Scott R, John H Esling and Lise Crevier-Buchman (2010). "A high-speed laryngoscopic investigation of aryepiglottic trilling a". In: *The Journal of the Acoustical Society of America* 127.3, pp. 1548–1558.

Mörner, M., F. Fransson and G. Fant (1963). "Voice register terminology and standard pitch." In: *STL-QPSR* 17.23.

Morris, Richard and Archie Bernard Harmon (2010). "Describing Voice Disorders". In: *Handbook of language and speech disorders*. Ed. by Jack Damico, Nicole Muller and Martin J. Ball. Chichester, U.K: Wiley-Blackwell, pp. 455–473.

Mundy, Rachel (2006). "Musical Evolution and the Making of Hierarchy". In: *World Of Music* 48.3.

Nettl, Bruno (2005). *The study of ethnomusicology: thirty-one issues and concepts*. Second. University of Illinois Press.

– (2006). "Response to Victor Grauer: On the Concept of Evolution in the History of Ethnomusicology". In: *World Of Music* 48.2.

Oates, J. (2009). "Auditory-Perceptual Evaluation of Disordered Voice Quality". In: *Folia Phoniatr Logop* 61, pp. 49–56.

O'Henry, Edward (1976). "The Variety of Music in a North Indian Village: Reassessing Cantometrics". In: *Ethnomusicology* 20.1, pp. 49–66. URL: http://www.jstor.org/stable/850820.

Olson, Lise (2001). "Some Personal Discoveries Regarding Vocal Use in Stage Combat." In: *Dal Vera, Rocco. The Voice in Violence and Other Contemporary Issues in Professional Voice and Speech Training Presented by the Voice and Speech Review. Cincinnati: Voice and Speech Trainers Association, Inc*, pp. 30–33.

Orr, R. et al. (2003). "An investigation of the parameters derived from the inverse filtering of flow and microphone signals". In: *Voice Quality: Functions, Analysis and Synthesis (VOQUAL '03)*. Taalwetenschap Otorhinolaryngology.

Osser, H. and F. Peng (1964). "A cross-cultural study of speech rate". In: *Language and Speech* 7, pp. 120–125.

Pearce, Eiluned, Jacques Launay and Robin IM Dunbar (2015). "The ice-breaker effect: singing mediates fast social bonding". In: *Royal Society Open Science* 2.10, p. 150221.

Pehlivan, Murat and İlter Denizoglu (2009). "Laryngoaltimeter: a new ambulatory device for laryngeal height control, preliminary results". In: *Journal of Voice* 23.5, pp. 529–538.

Pinker, Steven (1997). "How the mind works. 1997". In: *NY: Norton*.

Prather, Jonathan F et al. (2008). "Precise auditory-vocal mirroring in neurons for learned vocal communication". In: *Nature* 451.7176, p. 305.

Pressman, J. J. (1954). "Sphincters of the larynx". In: *A. M. A. Arch. Otolaryngol.* 59.2, pp. 221–236.

Proutskova, Polina and Michael Casey (2009). "You call THAT singing? Ensemble classification for multi-cultural collections of music recordings". In: *Proceedings of the International Symposium on Music Information Retrieval*.

Pulakka, Hannu (2005). "Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography". MA thesis. HELSINKI UNIVERSITY OF TECHNOLOGY, Department of Computer Science and Engineering.

Rahaim, Matthew (2006). "What Else Do We Say When We Say "Music Evolves"?" In: *World Of Music* 48.3.

Ramig, Lorraine Olson and Katherine Verdolini (1998). "Treatment efficacy: voice disorders". In: *Journal of Speech, Language, and Hearing Research* 41, pp. 101–116.

Riggs, Seth (1992). *Singing for the stars: a complete program for training your voice*. Alfred Music Publishing.

Rizzolatti, Giacomo and Laila Craighero (2004). "The mirror-neuron system". In: *Annu. Rev. Neurosci.* 27, pp. 169–192.

Rothenberg, M. (1973). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing". In: *The Journal of the Acoustical Society of America* 53, pp. 1632–1645.

– (1980). "Acoustic Interaction Between the Glottal Source and the Vocal Tract." In: *Vocal Fold Physiology*. Ed. by Kenneth N.Stevens and Minoru Hirano. Tokyo: University of Tokyo Press, pp. 305–328.

– (1981). "Research Aspects of Singing". In: *Publication issued by the Royal Swedish Academy of Music* 33, pp. 15–33.

Rouas, Jean-Luc and Leonidas Ioannidis (2016). "Automatic classification of phonation modes in singing voice: towards singing style characterisation and application to ethnomusicological recordings". In: *interspeech*. Vol. 2016, pp. 150–154.

Roubeau, Bernard, Nathalie Henrich and Michèle Castellengo (2009a). "Laryngeal vibratory mechanisms: The notion of vocal register revisited". In: *Journal of voice* 23.4, pp. 425–438.

– (2009b). "Laryngeal vibratory mechanisms: the notion of vocal register revisited". In: *Journal of Voice* 23.4, pp. 425–438.

Sadolin, Cathrine (2000). *Complete vocal technique*. Shout Publishing Copenhagen, Denmark.

Sakakibara, K et al. (2004a). "Growl voice in ethnic and pop styles". In: *Proc. Int. Symp. on Musical Acoustics*.

Sakakibara, Ken-Ichi et al. (2004b). "Physiological study of the supraglottal structure". In: *Proceedings of the International Conference on voice physiology and biomechanics*.

Savage, Patrick E. et al. (2015). "Statistical universals reveal the structures and functions of human music". In: *Proceedings of the National Academy of Sciences* 112.29, pp. 8987–8992. ISSN: 0027-8424. DOI: 10.1073/pnas.1414495112. eprint: http://www.pnas.org/content/112/29/8987.full.pdf. URL: http://www.pnas.org/content/112/29/8987.

Schutte, Harm K and Donald G Miller (1993). "Belting and pop, nonclassical approaches to the female middle voice: some preliminary considerations". In: *Journal of Voice* 7.2, pp. 142–150.

Shrivastav, Rahul and Christine M Sapienza (2006). "Some difference limens for the perception of breathiness a". In: *The Journal of the Acoustical Society of America* 120.1, pp. 416–423.

Soto-Morettini, D. (2006). *Popular Singing: A Practical Guide To: Pop, Jazz, Blues, Rock, Country and Gospel*. A&C Black. ISBN: 9780713672664. URL: http://books.google.co.uk/books?id=J91MyeNtdlkC.

Steinhauer, K and M McDonald Klimek (2017). "The Estill Voice Model: Theory & Translation". In: *San Francisco, CA: Estill Voice International*.

Steinhauer, Kimberly M, Deborah M Rekart and James Keaten (1992). "Nasality in modal speech and twang qualities: Physiologic, acoustic, and perceptual differences". In: *The Journal of the Acoustical Society of America* 92.4, pp. 2340–2340.

Stock, Jonathan P. J. (2006). "Clues from Our Present Peers?: A Response to Victor Grauer". In: *World Of Music* 48.2.

Stoller, Daniel, Simon Dixon et al. (2016). "Analysis and classification of phonation modes in singing". In:

Sundberg, J., P. Gramming and J LoVetri (1991). *Comparisons of Pharynx, Source, Formant and Pressure Characteristics in Operatic and Musical Theatre Singing*. Quarterly Progress and Status Report. Stockhom, Sweden: Speech Transmission Laboratory, Royal Institute of Technology.

Sundberg, J, I Titze and R Scherer (1993). "Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source". In: *J Voice* 7.1, pp. 15–29.

Sundberg, Johan (1987). *The science of the singing voice*. Illinois University Press.

– (2009). "Articulatory configuration and pitch in a classically trained soprano singer". In: *Journal of Voice* 23.5, pp. 546–551.

Sundberg, Johan and Margareta Thalén (2010). "What is "twang"?" In: *Journal of Voice* 24.6, pp. 654–660.

Sundberg, Johan et al. (1977). *The acoustics of the singing voice*. Scientific American.

Sundberg, Johan et al. (2004). "Estimating perceived phonatory pressedness in singing from flow glottograms". In: *J Voice* 18.1, pp. 56–62. DOI: 10.1016/j.jvoice.2003.05.006.

Sundberg, Johan et al. (2007). "Experimental findings on the nasal tract resonator in singing". In: *Journal of Voice* 21.2, pp. 127–137.

Svec, Jan G. and Svante Granqvist (2010). “Guidelines for Selecting Microphones for Human Voice Production Research”. In: *American Journal of Speech-Language Pathology* 19, pp. 356–368.

Svec, Jan G, Johan Sundberg and Stellan Hertegård (2008). “Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography”. In: *The Journal of the Acoustical Society of America* 123.1, pp. 347–353.

Takano, Sayoko, Kiyoshi Honda and Keisuke Kinoshita (2004). “Observation of cricothyroid joint motion using 3d high-resolution mri”. In: *Proc. Voice Phys. and Biomech.*

Thalén, M and J Sundberg (2001). “Describing different styles of singing: a comparison of a female singer’s voice source in "Classical", "Pop", "Jazz" and "Blues"”. In: *Logoped Phoniatr Vocol* 26.2, pp. 82–93.

Titze, I (2001). “Acoustic Interpretation of Resonant Voice”. In: *Journal of Voice* 15.4, pp. 519–528.

Titze, Ingo R (1994). *Principles of Voice Production*. Englewood Cliffs, New Jersey: Prentice Hall.

– (2006). “Voice training and therapy with a semi-occluded vocal tract: rationale and scientific underpinnings”. In: *Journal of Speech, Language, and Hearing Research* 49.2, pp. 448–459.

– (2008). “Nonlinear source-filter coupling in phonation: Theory”. In: *The Journal of the Acoustical Society of America* 123.5, pp. 2733–4.

Titze, Ingo R and Albert S Worley (2009). “Modeling source-filter interaction in belting and high-pitched operatic male singing”. In: *The Journal of the Acoustical Society of America* 126.3, pp. 1530–1540.

Trevarthen, Colwyn (1999-2000). “Musicality and the Intrinsic Motive Pulse: Evidence from Human Psychobiology and Infant Communication”. In: *Musicae Scientiae* Special Issue (Rhythm, Musical Narrative, and Origins of Human Communication), pp. 155–215.

Trundle, Deirdre (2005). *Changing voices: An approach to adolescent voice training*. Voicesource Publishing.

Vennard, W. (1967). *Singing: The Mechanism and the Technic*. Music Instruction Bks. C. Fischer. ISBN: 9780825800559. URL: `https://books.google.co.uk/books?id=nfgmgjqDwuMC`.

Walker, Jacqueline and Peter Murphy (2007). "A Review of Glottal Waveform Analysis". In: *PROGRESS IN NONLINEAR SPEECH PROCESSING*. Vol. 4391. Lecture Notes in Computer Science. Springer, pp. 1–21.

Wallin, Nils, Björn Merker and Steven Brown, eds. (2000). *The Origins of Music*. Cambridge, MA: MIT Press.

Watkins, Kate E, Antonio P Strafella and Tomáš Paus (2003). "Seeing and hearing speech excites the motor system involved in speech production". In: *Neuropsychologia* 41.8, pp. 989–994.

Weekly, Edrie Means and Jeannette L LoVetri (2009). "Follow-up contemporary commercial music (CCM) survey: who's teaching what in nonclassical music". In: *Journal of Voice* 23.3, pp. 367–375.

Wiggins, Geraint (2009). "Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music". In: *Proceedings of IEEE AdMIRe*.

Wilkins, Julia et al. (2018). "VocalSet: A Singing Voice Dataset". In: *19th International Society for Music Information Retrieval Conference Proc. (ISMIR2018)*.

Wilson, Stephen M et al. (2004). "Listening to speech activates motor areas involved in speech production". In: *Nature neuroscience* 7.7, p. 701.

Yanagisawa, E, S T Kmucha and J Estill (1990). "Role of the soft palate in laryngeal functions and selected voice qualities. Simultaneous velolaryngeal videoendoscopy". In: *The Annals of otology, rhinology, and laryngology* 99.1, pp. 18–28.

Yanagisawa, Eiji et al. (1989). "The contribution of aryepiglottic constriction to "ringing" voice quality—a videolaryngoscopic study with acoustic analysis". In: *Journal of Voice* 3.4, pp. 342–350.

Yanagisawa, Eiji et al. (1991). "Supraglottic contributions to pitch raising: videoendoscopic study with spectroanalysis". In: *Annals of Otology, Rhinology & Laryngology* 100.1, pp. 19–30.

Yoshinaga, Ikuyo and Jiangping Kong (2012). "Laryngeal vibratory behavior in traditional Noh singing". In: *Tsinghua Science and Technology* 17.1, pp. 94–103.

Yost, William A (1994). *Fundamentals of hearing: An introduction*. Academic Press.

Zeroual, Chakir, JH Esling and Lise Crevier-Buchman (2008). "The contribution of supraglottic laryngeal adjustments to voice: phonetic evidence from Arabic". In: *Logopedics Phoniatrics Vocology* 33.1, pp. 3–11.