# Breathy, Resonant, Pressed - Automatic Detection Of Phonation Mode From Audio Recordings of Singing

Polina Proutskova*, Christophe Rhodes*, Tim Crawford*, Geraint Wiggins**

June 18, 2013

*Department of Computing, Goldsmiths, University of London

** School of Electronic Engineering And Computer Science, Queen Mary University of London

In this paper we present an experiment on automatic detection of phonation modes from recordings of sustained sung vowels. We created an open dataset specifically for this experiment, containing recordings of nine vowels from multiple languages, sung by a female singer on all pitches in her vocal range in phonation modes breathy, neutral, flow (resonant) and pressed. The dataset is available under a Creative Commons license at `http://www.proutskova.de/phonation-modes/` .

First, glottal flow waveform is estimated via inverse filtering (IAIF) from audio recordings. Then six parameters of the glottal flow waveform are calculated. A 4-class Support Vector Machine classifier is constructed to separate these features into phonation mode classes. We automated the IAIF approach by computing the values of the input arguments – lip radiation and formant count – leading to the best-performing SVM classifiers (average classification accuracy over 60%), yielding a physical model for the articulation of the vowels.

We examine the steps needed to generalise and extend the experimental work presented in this paper in order to apply this method in ethnomusicological investigations.

## 1 Introduction

Phonation modes 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 used routinely by singing teachers to determine the progress of a student.

Johan Sundberg in his seminal work "The Science Of The Singing Voice" identifies four different phonation modes in singing: breathy, neutral, flow (called resonant by other authors) and pressed [Sundberg, 1987]. In this paper we present a method for automatic extraction of phonation modes from audio recordings of sustained vowels.

The introduction begins with a detailed discussion of Sundberg's definition of phonation modes. Then, examples of use of phonation modes in performance practice and in various disciplines are given. In particular, a possible ethnomusicological application of phonation modes is outlined: a falsification of the link between the singing style and the status of women in a society. We outline our method for supervised classification of phonation modes in Section 2. Section 3 describes the dataset which we have created in order to test this method, by means of the experiment presented and discussed in Section 4. We conclude with discussions and summaries in Section 5.

### 1.1 Phonation modes in singing: voice acoustic

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* (called *resonant* by other authors) 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 to four regions in the 2-D space spanned by glottal airflow and subglottal pressure (Figure 1.1). 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. The neutral mode corresponds to low airflow and low subglottal pressure, thus requiring the least physical effort. The flow phonation combines a high subglottal pressure and a high airflow.

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. 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. At the same time flow phonation enables the singer to gain a high sound level comparable to pressed phonation, which is significantly higher than in a neutral or a breathy mode. Also, flow phonation allows various resonances of the vocal tract to be used most effectively, while pressed phonation tends to restrict some of them and in breathy singing they are weak and obscured by non-harmonic parts of the spectrum.

This can be illustrated by means of the typical voice source signal waveforms. The graphs in Figure 1.2 are taken from Sundberg's book [1987, p. 85]; they show one full cycle of the vocal folds 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

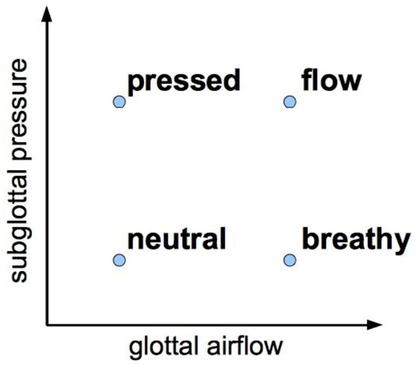


Figure 1.1: Phonation modes defined by Sundberg in his book "The Science of the Singing Voice" [Sundberg, 1987] schematically correspond to four regions in the 2-D space spanned by glottal airflow and subglottal pressure. E.g., when a large amount of air passes through the vocal folds by a low subglottal pressure, the resulting phonation mode is called breathy. In contrast, a high subglottal pressure with only a small amount of air passing the vocal folds gives rise to pressed phonation.

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. Flow phonation is described by Sundberg as the sweetspot where the maximal airflow is achieved retaining a closure of the vocal folds during the closed phase.

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, in which the physiology of sound production is fundamentally different.

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 Quotient (NAQ), the difference between the first and the second harmonics (H1-H2) and the closed quotient (ClQ). NAQ was

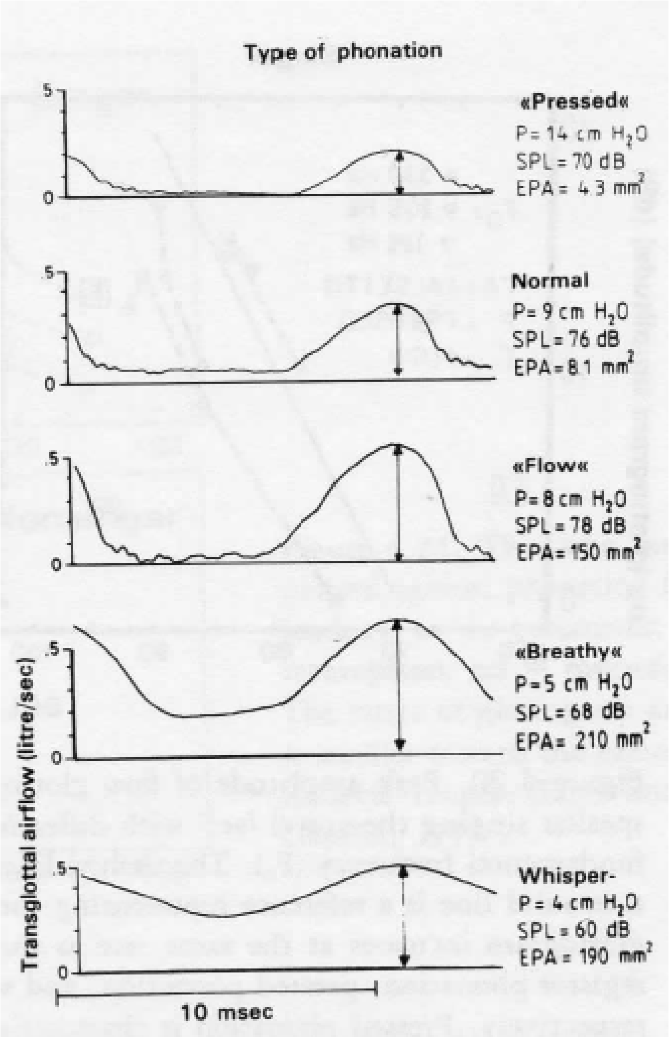


Figure 1.2: Typical graphs of the glottal flow waveform pulse functions in various phonation modes (from Sundberg 1987, p. 85, used with permission of Northern Illinois University Press)

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.

found to account for over 70% of the variations 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).

Like us 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.

## 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 skillfully 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 mode, such as Ella Fitzgerald's resonant, vibrating vocalising, e.g., on "Mack the Knife"[4], or Liza Minelli's "New York, New York"[5]; or from the pressed phonation, e.g., the tense, forceful voice of James Brown in "I feel good"[6].

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: ethnomusicology, singing education, medical research (phoniatrics, vocology), linguistics (phonetics) as well as in singing performance. The use

[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)

[4] http://www.youtube.com/watch?v=hRyDB4RWJdw (last accessed on 30/10/2012)

[5] http://www.youtube.com/watch?v=rgusCINe260 (last accessed on 30/10/2012)

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

of breathy, pressed or resonant singing production can be representative of an individual singing style as well as of a particular musical repertoire. In the experience of the first author as an ethnomusicologist and a singer, while each voice is different and two singers never sing the same way, repertoires within a music tradition (or sometimes across music traditions, e.g., Muslim call for prayer) display cultural preferences for the use of particular phonation mode(s), which are imposed on the singers performing in these repertoires. The first author's performance practice in a number of vocal traditions suggests, that in many cases flow phonation - the most economical way of producing loud and resonant sounds - is encouraged: 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 the disciplines of vocology (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.

## 1.3 Ethnomusicological motivation

In this section, we describe one possible application of phonation modes and their automated detection in a wider interdisciplinary investigation.

Ethnomusicologist Alan Lomax denotes the dichotomy between tense and relaxed vocalisation as the *vocal tension* parameter. In his Cantometrics textbook [Lomax, 1976] he describes relaxed singing as mellow, wide and richly resonant, while tense voices sound narrow, pinched and restricted in resonance (p. 125). In a large-scale experiment carried out on over 5000 audio clips representing more than 500 musical traditions, Lomax and his team sought for correlations between the singing style prevalent in a society and its societal traits (such as stratification or attitude to child rearing). The idea behind the Cantometrics project was that singing, being a mode of communication which is highly regulated and encapsulating peoples values and traditions, must reflect general communication patterns typical for the given society. A parametrization system of 36 singing

| Musical factor | Societal descriptor | # cultures | p-value |
|---|---|---|---|
| Differentiation (information load) | Productive scale | 157 | .001 |
| Ornamentation | Large domesticated animals | 97 | .001 |
| Orchestral organisation | States | 82 | .001 |
| Cohesiveness | Community solidarity | 143 | .001 |
| Choral organisation | Solidarity | 102 | .01 |
| Noise/Tension in voice production | Severity of sex sanctions | 117 | .001 |
| Energy level (volume, accent, pitch) | Extra-local government hierarchies | 151 | .001 |
| Irregular to regular rhythm | Infant/child indulgence | 40 | .001/.01 |
| Melody (complex/simple) | Large/small settlement | 124 | .001 |

Table 1: Correlations between musical and societal parameters discovered in Cantometrics. This table gives an overview of general relationships for groups of Cantometric parameters (factors). For more details on correlation see Lomax [1976], pp. 22-28 and 260-269

style descriptors was laid out and statistical analysis of correlations between those and the anthropological data was performed. The Cantometrics team found that in societies where narrow, squeezed, tense vocalisation is the norm, pre-marital sex is strongly forbidden for women and vice versa, where singing is relaxed and open-throated, the rules regarding the pre-marital behaviour of women are also more relaxed. A hypothesis about the relationship between vocal tension and subordination of women was put forward.

Though a consistent and statistically sound picture of relationships between musical style and societal traits was reported by the Cantometrics team (see Table 1), the project was widely criticised by ethnomusicologists (see, e.g., Nettl, 2008 for a summary of a discussion between Lomax and Herndon). The main critique points were: the subjectivity of human ratings of Cantometrics musical parameters; that the musical examples chosen by Lomax do not represent the diversity within the given cultures adequately [O'Henry, 1976]; that correlations between the production mode and the musical organisation found by Lomax do not hold for some societies [Feld, 1984]; that focusing exclusively on musical surface and leaving out the cultural context and practices leads to superficialities [Feld, 1984]. The authors would add to this list inconsistencies in the definition of vocal tension (tense vocalisation can be produced with a wide, open throat) as well as in the training examples for the parameter raters. At the same time major thinkers in ethnomusicology acknowledge the value of Cantometrics and the unique and unsurpassed scope of the project [Nettl, 2005].

No systematic verification of Cantometrics methodologies and results has ever been attempted. This task has been especially unapproachable, because the data which was used in the Cantometrics project and its complete methodology and results have never been published.

Recently, a new discussion involving Cantometrics has emerged. In a truly vast attempt to outline his views of the global history of human musical style, its origins and evolution, Victor Grauer [2006b] (who was Lomax's assistant and the co-inventor of Cantometrics in 1960s) relies heavily on Cantometric 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, 2006a; Rahaim, 2006; Cross, 2006; Mundy, 2006].

Phonation modes might provide a suitable physiological/acoustical model for the perceptual categories (tense, narrow vs. wide, relaxed) applied by Lomax to define vocal tension; they might be used to reformulate the vocal tension definition to make it more objective and measurable. Automatic extraction of phonation modes from recordings of singing could then provide the basis for a new approach to a revision of the Cantometrics experiment. It would help to address the main methodological weaknesses of the Cantometrics approach: the subjectivity of human ratings as well as the limited representation of each culture in the sample. With automatic phonation mode extraction, any number of new musical samples could be included in the experiment revision without the need for a subjective and labour-intensive human rating procedure.

## 2 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. 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 2.2.

## 2.1 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. Thus the standard spectral features such as MFCCs and chroma are not well suited for the task. 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 2.1). 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.

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 et al. 2012, Gudnason et al. 2012]. The vocal tract (throat, mouth and in some cases nose) forms the tube, which is characterized 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 2.1).

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

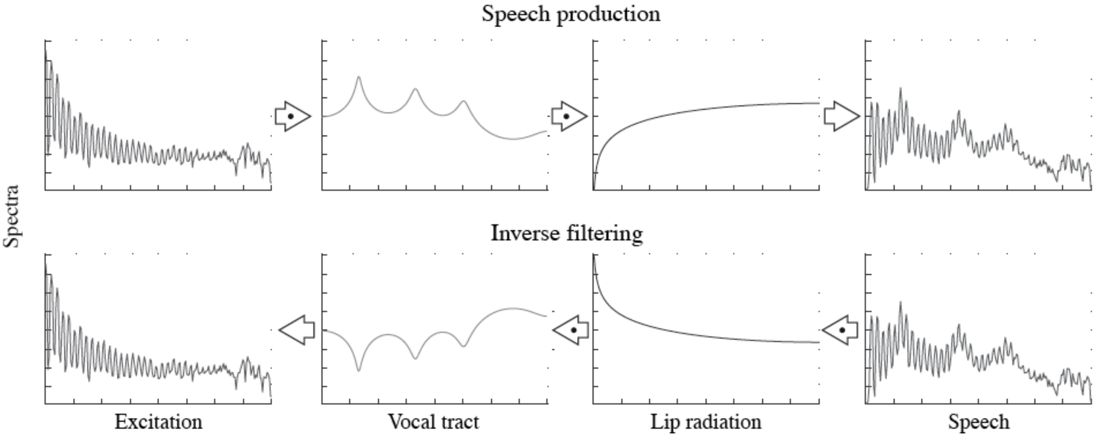


Figure 2.1: Inverse filtering: the upper row represents the separated speech production model; the lower row illustrates the corresponding inverse filtering process, in which the lip radiation and vocal tract filters are inverted to acquire an estimate for the glottal flow waveform (reproduced from Airas 2008, p. 50, with permission by Informa Group)

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.

## 2.2 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 publically-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 optimization 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 flow and 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 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 at `http://www.proutskova.de/phonation-modes/` .

## 3.1 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 2. These sounds were sung on all pitches on a semitone scale from A3 to G5, in every phonation mode given in Table 3. The phonation modes correspond to Sundberg's definitions of breathy, neutral, flow and pressed phonation [Sundberg, 1987], (see Section 1.1).

## 3.2 The singer

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.

| 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 *т**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**т*, somewhat similar to English *c**au**ght* | O |
| [u:] | /u/ - high back rounded, as in English *b**oo**t*, German *F**u**ß*, Russian *пл**у**г* | 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**том*, similar to [æ] in English *c**a**t* | AE |
| [ɨ:] | High central unrounded vowel, Russian /ы/ as in *т**ы***, similar to English *ros**e**s* | Y |

Table 2: The vowels represented in the 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 |

Table 3: The range in which a given vowel in a given phonation mode is represented in the dataset. Flow and pressed phonation could only be realised up to the upper part of the modal register (H4-C4).

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. The singer's break between the modal and the falsetto register is around E5, thus the surrounding pitches (D5 to F5) can also be less reliable. Still we decided to include vocalisation in the falsetto register into the dataset to make it more representative, thus all pitches up to G5 were included.

In the head voice the singer was unable to produce flow and pressed sounds, thus these phonation modes are only represented up to the upper range of the modal register (see Table 3). Above C5 it becomes impossible to sing most vowels in the flow and the pressed modes; at the same time, the neutral mode in the middle and head voice partly gains the qualities of the flow mode, such as intensity and richness in overtones, though it is very different from the chesty flow phonation. The singer reported from her experience of teaching Russian traditional singing, which heavily uses flow phonation, that this limit is typical for female singers.

Why this is the case seems to be an unsolved problem. This seems to be common among singers of various traditions in Europe and the Near East, but it is unclear whether in other cultures (e.g. in East Asian traditions) the singers are in fact capable of producing flow and pressed vocalisation in their head register. This observation leads to the question whether the ability to use particular phonation modes on particular pitches is innate or ontogenetic (learned through culture).

While the singer was confident in breathy, neutral and flow phonation, pressed vowels seemed to present problems and were mostly exaggerated to the point where they were considerably uncomfortable and unhealthy to produce. The reason for this is the fact that the singer routinely used breathy, neutral and flow phonation in her singing practice while pressed phonation was only used marginally.

In the lower range, at G3 and below, the neutral phonation becomes more chesty and therefore quite similar to the flow mode—for this reason recordings below A3 were excluded from the dataset.

The singer apparently had more difficulties with some vowels than with others in particular modes. For example, high front sounds like [i:] and [y:] proved to be harder to achieve in flow phonation. The vocal results for the sounds [y:] and [ɨ:] in pressed phonation on the highest pitches of the chest register were unstable and were not included in the dataset.

## 3.3 Recording conditions

The recordings were made with a professional dynamic microphone from Electro-Voice, model no. N/D375A. The model was chosen because of its flat response: $+10dB \pm 1dB$ between 200 Hz and 15000 Hz (Figure 3.1). 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 3.1 was measured. Svec and Gramqvist [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

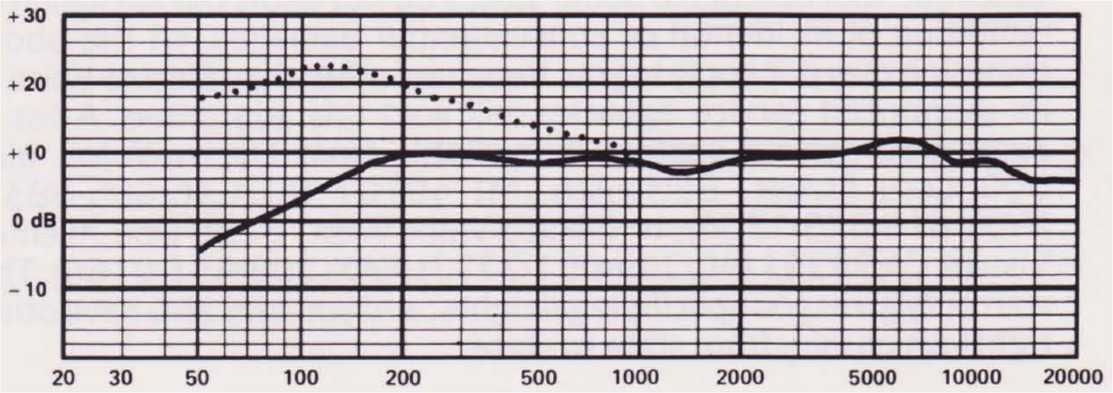


Figure 3.1: N/D357A microphone frequency response curve (thick curve). The thin curve marks the proximity effect which only takes effect at the distance of 12 inches (30 cm) or closer.

pre-amplifier and audio interface from M-Audio. It was then connected to a MacBook Pro 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) [IAS, 2009]. The recording session took place in a quiet room environment. The requirement of a signal-to-noise ratio of at least 15 dB has been adhered to [Svec and Granqvist, 2010].

### 3.4 The dataset availability

The dataset is available for download at `http://www.proutskova.de/phonation-modes/` under Creative Commons CC BY-NC-SA license. This license allows free sharing of the dataset as 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.

Further content additions and future support for the dataset are planned. Also, additions from other parties will be welcome.

## 4 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 2.

## 4.1 Experiment design

Because of the model constraints outlined in Section 2.2 the experiment was performed separately for each vowel. The flow chart of the experiment is given in Figure 4.1. 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 flow mode and also the pressed mode for vowels 'UE' and 'Y', which are represented at all pitches except C5).

For feature extraction we used an implementation of the IAIF algorithm (see Section 2.1) 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 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

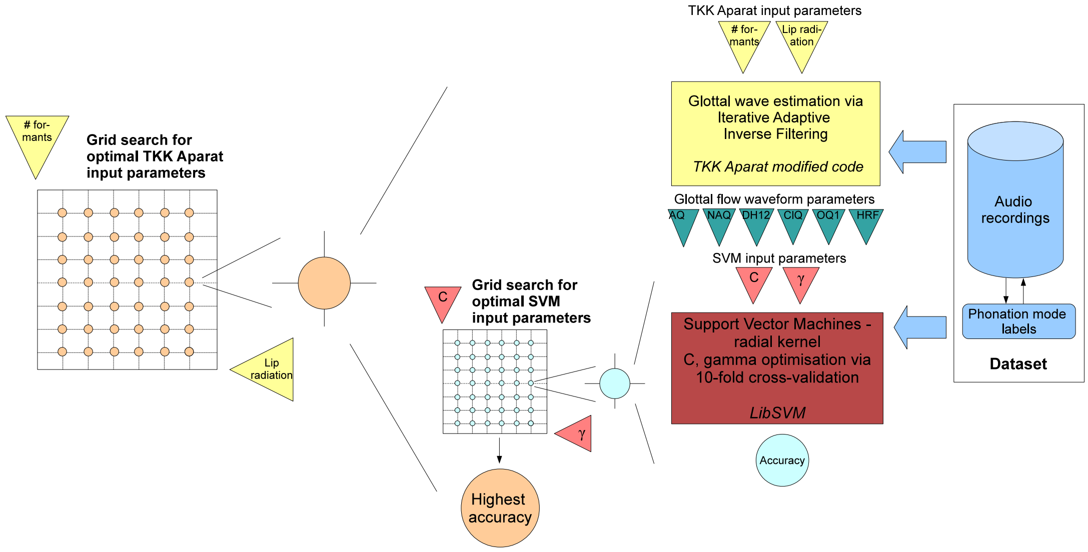

TKK Aparat input parameters
# for-mants
Lip radi-ation
Glottal wave estimation via Iterative Adaptive Inverse Filtering
TKK Aparat modified code
Glottal flow waveform parameters
AQ
NAQ
DH12
CIQ
OQ1
HRF
SVM input parameters
C
γ
Support Vector Machines - radial kernel
C, gamma optimisation via 10-fold cross-validation
LibSVM
Accuracy
Audio recordings
Phonation mode labels
Dataset
# for-mants
Grid search for optimal TKK Aparat input parameters
Lip radiation
C
Grid search for optimal SVM input parameters
γ
Highest accuracy

Figure 4.1: 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.

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 utilasing the six low-level features returned by IAIF and the phonation mode labels from the dataset is performed; 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.

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 4.2 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.

## 4.2 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 4.3). 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.

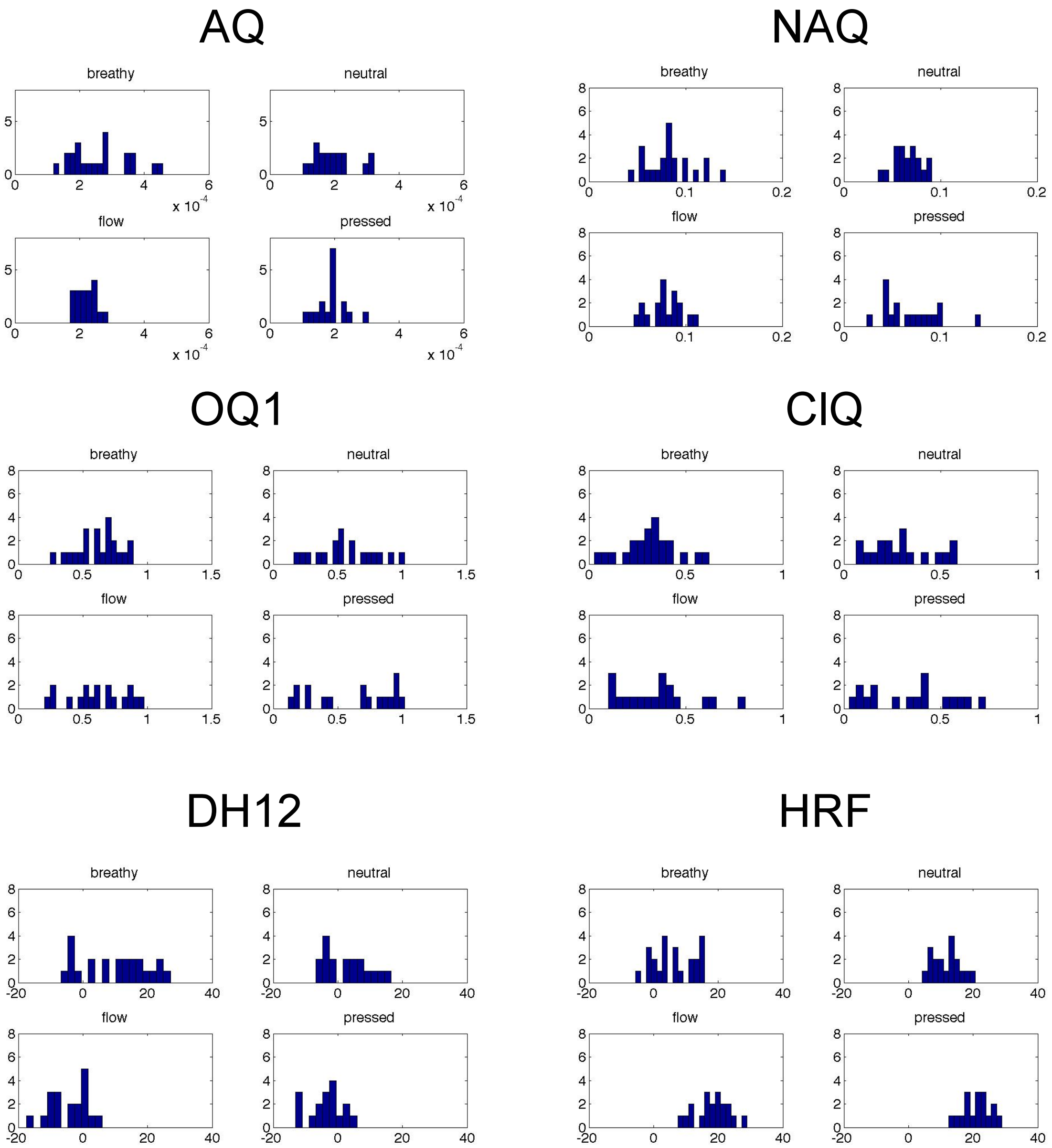


Figure 4.2: The distributions of the six voice source waveform descriptors for each phonation mode for the vowel A.

| 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 |

Table 4: 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 accurace 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.

Fine grid search results with the corresponding optimal values of input parameters are given in Table 4. The average accurace 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 4 also gives the values of the input parameters leading to the highest phonation mode classification accuracy. These optimal values for lip radiation and number of vocal tract segments - the input parameters of the IAIF algorithm / TKK Aparat implementation - are plottet together in Figure 4.4 to allow comparison. Confusion matrices for classification with the optimal input parameters are given in Figure 4.5.

## 4.3 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 2.2.

Figure 4.4 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 physical properties of the tested vowels - the opening of the mouth and the complexity of the form of the vocal tract.

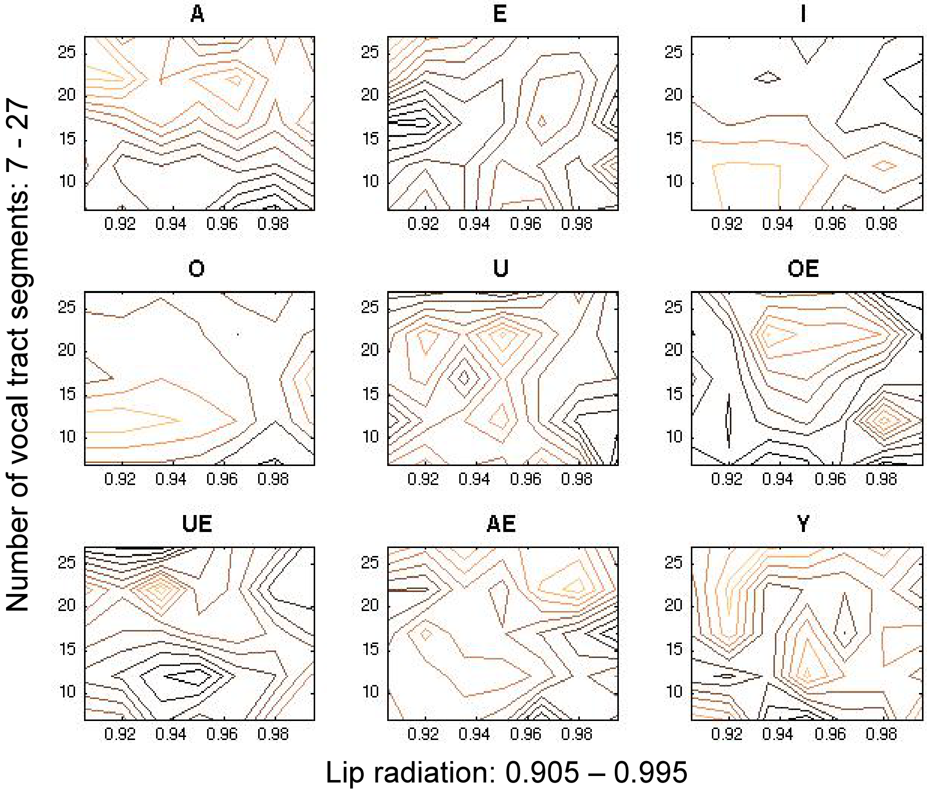


Figure 4.3: Coarse grid search results. The graphs represent phonation mode classification accuracy as a function of lip radiation (x axes) and number of vocal tract segments (y axes). Number of vocal tract segments was iterated from 7 to 27 in 5-steps; lip radiation from 0.905 to 0.995 in steps of 0.015. 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 2.2. Optimal solutions were picked manually taking into account besides the accuracy values illustrated here also the stability of the solution (expressed in standard deviation) as well as the simplicity of the model.

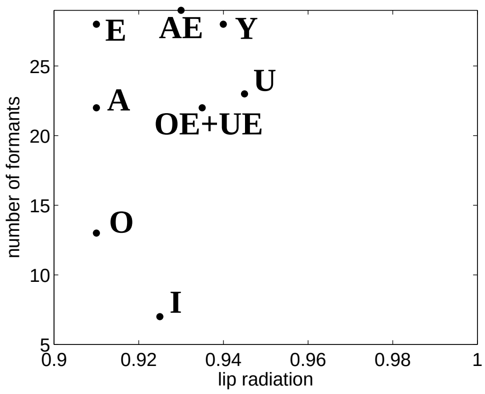
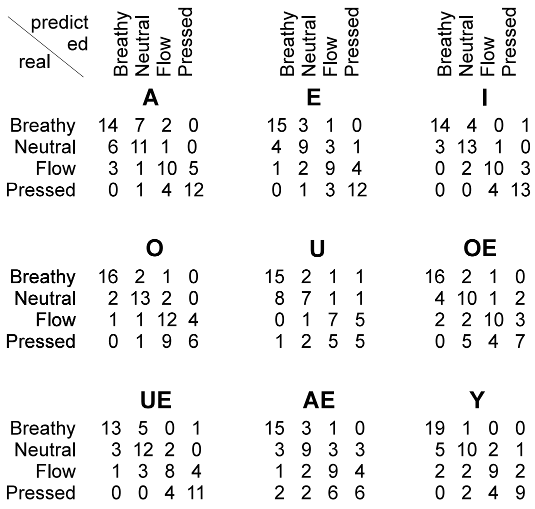


Figure 4.4: Optimal solutions for all vowels. For each of the nine vowels, the values of the IAIF input arguments - lip radiation and number of vocal tract segments - leading to the highest phonation mode classification accuracy have been plotted in one space to allow comparison. The optimal lip radiation 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.



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

Figure 4.5: 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 as well as between flow and pressed modes, and less confusion across these two groups. 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 physical 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 4.4.

The issue of acquiring more reliable ground truth will have to be considered. Currently the labels in the dataset are based on the first author'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 determined 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. The singers we have worked with can sing in flow and pressed phonation up to a fifth below the break, some can go as high as up to a third (see our discussion of this point in the dataset description 3.2). In Western classical singing school flow phonation is used extensively in the chest voice, while in the head voice and 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 flow phonation is retained in the mix of chest and head voice, resulting in a loud and tense vocalisation. It remains an open question whether these limits for the flow and pressed sound production are culturally specific or result from general human physiology.

## 4.4 Automatic vocal tension estimation: how to get there

As we mentioned in the introduction, one of the motivations for us to take up research on phonation modes has been the interest in the subordination of women hypothesis put forward by Alan Lomax in the course of the Cantometrics experiment (Section 1.3). Though our work on automatic extration of phonation modes from controlled recordings of sustained vowels is valuable in its own right and opens up many avenues for futher research, we would like to discuss the revision of Cantometrics in more detail here.

The Cantometrics project was a complex comparative music study of an unsurpassed scope, whose potential has not been explored so far. The reason for this is that its results have not been independently validated and that its methodology displays weaknesses that have not been addressed (see Section 1.3 for more details). Automatic extraction of phonation modes from audio by means of MIR methods might provide the basis for a new approach to assessing one of the most speculative, controversial and political outcomes of the Cantometrics experiment - the link between the tense, narrow singing style and the subordination of women in a society.

The results presented in this paper were achieved on a set of specially produced recordings of sustained vowels. This is far away from automatic determination of prevalent phonation modes for real-life, multipart singing recordings as they are present, e.g., in the Cantometrics dataset. This subsection discusses the steps necessary to reach the general solution.

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 20th Century. Cantometrics measures this variation along 36 musical style descriptors (see Section 1.3). 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 origin 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 by one of the most general MIR tasks: automatic extraction of a high-level descriptor from a highly heterogeneous audio dataset. We suggest to approach it by solving incremental problems. We start with controlled conditions, where variation is introduced in pitch and in phonation only. There is just one singer, thus no inter-singer variation is present; recording conditions are the same for all tracks; the singer sings only sustained vowels, so non-harmonic sounds like consonants do not interfere. This is the setting for the experiment presented in this paper.

The articulation of the vowel sound is crucial for the model used for feature extraction in our approach. This means that the information about the vowel cannot be discarded or automatically extracted from audio by the same method. Therefore, for a generalisation of the presented experiment to real-life recordings other methods for vowel determination will have to be used prior to feature extraction. The next step of generalisation could thus be based on real-life recordings of the same singer and employ a method for automatic vocal sample extraction and vowel determination. This experiment could also look at the role of recording conditions in accurate phonation mode extraction.

In the following step recordings by more than one singer could be analysed and the inter-singer variation studied. This can be further extended by including recordings by singers from different cultures employing a variety of singing techniques. After that group singing could be investigated, including various group mixtures and singing styles. Accompanied singing will certainly present a problem: a method for automatic extraction of relevant samples in which the voices are unaccompanied or highly dominant will have to be introduced. At every stage the influence of recording quality and conditions will have to be investigated and taken into account.

Given that all the incremental stages described above succeed, the most general task of automatic extraction of phonation modes from a dataset as varied as the Cantometrics collection can be addressed.

# 5 Conclusions

Phonation mode is an important characteristic of singing, playing a vital role in many singing-related disciplines. It remains under-researched, one of the reasons being the lack of reference and training data. Our paper closes this gap, approaching it in three directions: we have created a publicly available dataset, we present initial classification results and we place our current work into a wider interdisciplinary context, suggesting further research directions.

### The dataset

We have recorded, annotated and provided online access to the first systematic dataset on phonation modes in singing. It contains recordings of nine sustained vowels produced by one female singer. Pitches from almost two octaves are covered, including the register break. Recoding conditions were controlled and documented.

The dataset is available online under a clear license, making it easy for other researchers to validate our findings and to suggest other strategies, whose performances can then be compared on the basis of the dataset. This makes research more transparent and facilitates competition as well as collaboration.

### The initial classification results

In this paper we present the first experiment on automatic extraction of phonation modes from audio recordings of sustained vowels, based on the dataset we created and made available publicly. This is the first experiment that systematically includes all pitches and phonation modes for nine vowels from three languages in the training set and calculates the highest classification accuracy via batch processing. IAIF approach is automated and optimal values for its input parameters for all vowels are reported.

The results of this initial classification experiment being above 60% accuracy for a four-class classifier demonstrate that there is structure in the data and that further work on improving classification is justified. The dataset together with the experiment's results provide a benchmark for future research on automatic classification of phonation modes from audio recordings of singing.

### Interdisciplinary context

We offer a discussion of a large-scale ethnomusicological investigation, that would answer an open question in the discipline: whether cultural preferences for particular phonation modes in singing are related to the status of women in a society. We believe it important to build bridges between humanistic and scientific research and to study human behaviour by means of quantitative methods. With the current work we have made the first step on a long road to these goals. Our study was motivated by a humanistic question, as opposed to the bottow-up approach mainly practiced in the MIR today, in which the available technology defines the sorts of problems to be addressed. In the understanding that the road we have chosen is indeed long and unpredictable, we have made the best effort to make it as easy as possible for other researchers to pick up the thread.

## 6 Acknowledgements

We would like to sincerely thank Victor Grauer, the co-inventor of Cantometrics, for suggesting to focus on the subordination of women hypothesis, which gave us the inspiration and the motivation for the current work.

We are grateful to Johan Sundberg for his recommendations, e.g., on the recordings set up and his general support.

Our special thanks go to the peer reviewers for their balanced, insightful and fair reviews, which helped us to substantially revise the text.

## References

Matti Airas. TKK aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33:49–64, 2008.

P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.*, 11:109–118, 1992.

D. Zangger Borch and Johan Sundberg. Some phonatory and resonatory characteristics of the rock, pop, soul, and swedish dance band styles of singing. *J Voice*, 25(5):532–7, Sep 2011. doi: 10.1016/j.jvoice.2010.07.014.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

DG Childers and C-F. Wong. Measuring and modeling vocal source-tract interaction. In *IEEE Trans Biomed Eng.*, volume 41, 1994.

Peter Cooke. Response to echoes of our forgotten ancestors. *World Of Music*, 48(2), 2006.

Ian Cross. Four issues in the study of music evolution. *World Of Music*, 48(3), 2006.

T. Drugman, T. Dubuisson, A. Moinet, N. D'Alessandro, and T. Dutoit. Glottal source estimation robustness. In *Proc. of the IEEE International Conference on Signal Processing and Multimedia Applications (SIGMAP08)*, 2008.

Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech and Language*, 26:20–34, 2012.

G. Fant. *Acoustic theory of speech production.* Mouton, The Hague, Netherlands, 1960.

Steven Feld. Sound structure as social structure. *Ethnomusicology*, 28(3):383–409, 1984. ISSN 00141836. URL `http://www.jstor.org/stable/851232`.

Bjorn Fritzell. Inverse filtering. *Journal of Voice*, 6(2):111–114, 1992.

Emil Froeschels. Hygiene of the voice. *Arch Otolaryngol.*, 38(2):122–130, 1943.

Svante Granqvist. *Computer methods for voice analysis.* PhD thesis, Department of Speech, Music and Hearing, Stockholm, 2003.

Victor A. Grauer. Echoes of our forgotten ancestors: Some points of clarification. *The World Of Music*, 48(3), 2006a.

Victor A. Grauer. Echoes of our forgotten ancestors. *The World Of Music*, 48(2), 2006b.

Victor A. Grauer. New perspectives on the kalahari debate: a tale of two 'genomes'. *Before Farming, the archaeology and anthropology of hunter-gatherers*, 2, 2007.

Jon Gudnason, Daniel P.W. Ellis Mark R.P. Thomas, and Patrick A. Naylor. Data-driven voice source waveform analysis and synthesis. *Speech Communication*, 54:199–211, 2012.

David M. Howard. Electrolaryngographically revealed aspects of the voice source in singing. *Logopedics Phoniatrics Vocology*, 35(2):81–89, 2010.

*Guidelines on the Production and Preservation of Digital Audio Objects: Standards, Recommended Practices and Strategies (IASA-TC 04).* IASA (International Association for Sound- and Audiovisual Archives) Technical Committee, 2 edition, 2009.

Laura Lehto, Matti Airas, Eva Björkner, Johan Sundberg, and Paavo Alku. Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *J Voice*, 21(2):138–50, Mar 2007. doi: 10.1016/j.jvoice.2005.10.007.

Alan Lomax. *Cantometrics: An Approach To The Anthropology Of Music.* Number 94720. The University of California, Extension Media Center, Berkeley, California, 1976. accompanied by 7 cassettes.

Rachel Mundy. Musical evolution and the making of hierarchy. *World Of Music*, 48(3), 2006.

Bruno Nettl. *The study of ethnomusicology: thirty-one issues and concepts.* University of Illinois Press, second edition, 2005.

Bruno Nettl. Response to victor grauer: On the concept of evolution in the history of ethnomusicology. *World Of Music*, 48(2), 2006.

Bruno Nettl. Comparative study and comparative musicology: comments on disciplinary history. In Albrecht Schneider, editor, *Systematic and Comparative Musicology: Concepts, Methods, Findings*, volume 24 of *Hamburger Jahrbuch für Musikwissenschaft*, pages 295–314. Peter Lang, 2008.

Edward O'Henry. The variety of music in a north indian village: Reassessing cantometrics. *Ethnomusicology*, 20(1):49–66, 1976. URL `http://www.jstor.org/stable/850820`.

R. Orr, B. Cranen, F.I.C.R.S. de Jong, C. d'Alessandro, and K.R. Scherer. 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, 2003.

Hannu Pulakka. Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. Master's thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, Department of Computer Science and Engineering, February 2005.

Matthew Rahaim. What else do we say when we say "music evolves"? *World Of Music*, 48(3), 2006.

Lorraine Olson Ramig and Katherine Verdolini. Journal of speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, 41:101–116, February 1998.

M. Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *The Journal of the Acoustical Society of America*, 53:1632–1645, 1973.

M. Rothenberg. Acoustic interaction between the glottal source and the vocal tract. In Kenneth N.Stevens and Minoru Hirano, editors, *Vocal Fold Physiology*, pages 305–328. University of Tokyo Press, Tokyo, 1980.

Jonathan P. J. Stock. Clues from our present peers?: A response to victor grauer. *World Of Music*, 48(2), 2006.

Johan Sundberg. *The science of the singing voice*. Illinois University Press, 1987.

Johan Sundberg, Margareta Thalén, Paavo Alku, and Erkki Vilkman. Estimating perceived phonatory pressedness in singing from flow glottograms. *J Voice*, 18(1):56–62, Mar 2004. doi: 10.1016/j.jvoice.2003.05.006.

Jan G. Svec and Svante Granqvist. Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, 19:356–368, 2010.

M Thalén and J Sundberg. Describing different styles of singing: a comparison of a female singer's voice source in "classical", "pop", "jazz" and "blues". *Logoped Phoniatr Vocol*, 26(2):82–93, 2001.

Jacqueline Walker and Peter Murphy. A review of glottal waveform analysis. In *PROGRESS IN NONLINEAR SPEECH PROCESSING*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer, 2007.