Master thesis on Sound and Music Computing
Universitat Pompeu Fabra

Analysis and Automatic Classification of Phonation Modes in Singing

Furkan Yesiler

Supervisor: Rafael Ramirez

September 2018

Universitat
Pompeu Fabra
Barcelona

Master thesis on Sound and Music Computing
Universitat Pompeu Fabra

Analysis and Automatic Classification of Phonation Modes in Singing

Furkan Yesiler

Supervisor: Rafael Ramirez

September 2018

Universitat
Pompeu Fabra
Barcelona

Contents

Acknowledgements

First and foremost, I would like to thank my supervisor, Rafael Ramirez, for his advice and support over the course of my research. Thank you for your confidence in me, and your encouragement throughout the year.

In addition, I would like to thank Baris Bozkurt for his help and valuable guidance during this year. I am particularly grateful for his advice that led me applying to this program and being a part of this community.

Also, I would like to express my gratitude to Xavier Serra and SMC Committee for providing me the opportunity to be a part of SMC and MTG.

Thanks to all my friends from SMC for the time we've spent together. Hope to see you all throughout the upcoming years.

Finally, I sincerely thank all my family and Sonja for their encouragement and unending moral support. This accomplishment would not have been possible without them. Thank you.

Abstract

Analysis of expression in singing voice is gaining more importance as the current assessment systems fail to consider important resources in expressive singing, e.g. phonation modes. Phonation modes have been divided into four categories (breathy, pressed, neutral and flow) that correspond to levels of glottal adduction force. This thesis focuses on the analysis and automatic classification of phonation modes, and proposes a visual feedback system designed for singing voice assessment, vocal education and musicological analysis.

We propose to use a wide range of audio descriptors in order to extract information from the audio signal and to perform feature selection for reducing the dimension of the feature set. A supervised classification approach is applied with making use of Multi-Layer Perceptrons (MLP). The hyperparameters of the model are optimized with cross validation on training subsets. The results of the evaluation of the obtained model outperform the state of the art methods.

In order to generalize the feature analysis to avoid bias caused by having insufficient data we curated two new datasets for phonation modes research. Finally, the designed visual feedback system is tested with singing students and teachers to assess its usefulness for educational purposes.

Keywords: Singing Voice, Phonation Modes, Visual Feedback System, Automatic Classification

Chapter 1

Introduction

In our first chapter, we present an introduction to our research. We begin with discussing expressivity in music and singing voice, and give a description of phonation modes in singing. Then, we summarize previous works on this subject, and explain the goals and contributions of this thesis.

Context

In this section, we present a brief summary of expression in music performances and in singing. The concept of phonation modes is considered as an expressive resource of singing voice [1], and to understand the expressivity in music/singing voice, here, we present a short description based on previous works.

Expression in Music Performance

Expression is an essential part of a music performance and often expected from musicians [2]. Form and structure of a musical piece that are specified by a composer through a music score [3] can be performed by musicians in a number of acceptable ways [4] with making use of musically expressive resources [5]. The role of expression is emphasized by Oxford English Dictionary [6] by defining music as "that one of the fine arts which is concerned with the combination of sounds with a view to beauty of form and the expression of emotion."

There have been many attempts to define music expressivity. Palmer and Hutchins [7] proposed the following definition,

> Performers add variation to music; they manipulate the sound properties, including frequency (pitch), time, amplitude, and timbre (harmonic spectrum) above and beyond the pitch and duration categories that are determined by composers. These manipulations are called 'musical expression'.

Based on the aforementioned definition, we can categorize the facets of music expressivity into four groups: pitch, timing, intensity and timbre. These expressive dimensions can be manipulated independently while it is also a common practice to combine them to express certain styles and emotions.

An important point for music expressivity is that these manipulations are expected to be controlled and lie within certain boundaries. Sundberg [8] states that "Listening to a good performance can be as exciting as it is agonising to listen to a performance with a neutral or an inappropriate expression."

Juslin and Sloboda [9] stated that music expressivity can convey emotions, and a comprehensive review of the relationship between music expressivity and emotions was presented by Justin and Laukka [2]. They reported that high pitch variability is associated with anger and happiness while low values are used for fear and sadness. Performers use fast tempo for anger, fear and happiness, and slow tempo is commonly used to express sadness and tenderness. High values of loudness can be observed for anger, and low values are most commonly seen for fear and sadness.

Expression in Singing Voice

The singing voice is considered by many to be the most expressive instrument. Darwin [10] stated that "With many kinds of animals, man included, the vocal organs are efficient in the highest degree as a means of expression." Humans use their voice for many communicative purposes, including conveying emotions. Scherer [11] stated that "If the origin of music is indeed to be sought in the emotional expressions of the human voice, it should be human vocal music–singing that should be most prone to evoke strong emotional feelings in the listener." When singing voice is used to convey a certain emotion, changes are observed in respiration, phonation and articulation, and these changes affect the parameters of the acoustic signal. As a result of the modifications in the acoustic signal, singers can control expressivity in pitch, timing, intensity and timbre.

Manipulations of pitch is a well studied aspect of music/vocal expressivity. Perceived pitch is acoustically related to the fundamental frequency (F0), and in the literature, those two terms are often used to refer to F0. Performers use alterations of pitch with many types of instruments, e.g. string, woodwind, singing voice, for expressive purposes. Most common expressive devices regarding pitch alterations using singing voice are intonation and vibrato.

Temporal aspects of music/vocal expressivity include tempo and timing. Tempo indicates the overall pace of a musical segment and is commonly represented by the number of beats per minute. Timing, on the other hand, allows performers to express deviations from the predefined musical notation.

Intensity is related to the perceived loudness of a musical piece and defined by Juslin and Laukka [2] as a "measure of energy in the acoustic signal." Changing intensity is a commonly used expressive device to reflect certain styles.

Timbre, compared to the other dimensions, is the least studied aspect of music/vocal expressivity. Although it can be considered as a characteristic feature of instruments, different timbrel characteristics can be realized for the same instrument/singing voice using different techniques. In this study, we discuss the properties of phonation modes which we consider as a timbrel aspect of vocal expressivity.

Phonation Modes

Before continuing with the explanation of phonation modes, here, we provide a description of the voice organ and how it works. For a detailed description on the topic, we refer the reader to “The Science of the Singing Voice” by Sundberg [12].

Singing Voice Production

The voice organ consists of three systems [12] as can be seen in Figure 1: the breathing apparatus, the vocal folds and the vocal tract. The breathing apparatus is sending air to the vocal folds, and as a result of this airstream, the vocal folds may be brought together by the adduction movement or separated by the abduction movement. Vibration rate of the vocal folds determine the phonation frequency or the perceived pitch. The produced sound in vocal folds passes the vocal tract which acts as a resonator.

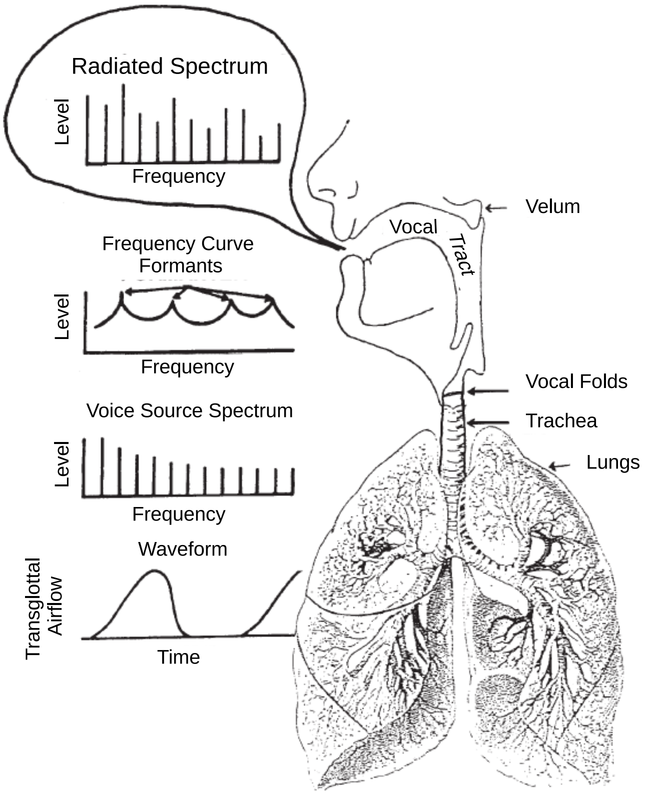

Figure 1: Summarized illustration of the vocal organ and voice production (Figure taken from [13], used with the permission of the author)

The role of the breathing apparatus (lungs) is to provide an overpressure of air for sound production. This overpressure (subglottal pressure) can be controlled in order to control loudness: a higher subglottal pressure results in a louder sound. In speech, an increase in subglottal pressure, or loudness, is associated with an increase in pitch [14] while for singing, phonation frequency and loudness are considered as two independent phonatory parameters [15].

The air pressure on the vocal folds affects the characteristics of the voice source. Vocal folds vibrate at a certain frequency, and in each cycle, closed and open phases of the glottis can be observed. During the closed phase, the vocal folds are adducted for a complete or a partial vocal fold closure while during the open phase, an air pulse passes through the glottis. The point where the maximum amount of air passes is referred as the peak flow. The duration of the closed phase and the amount of air passing through the glottis show variability depending on phonation modes which we explain in detail in the next subsection.

The vocal tract acts as a filter for the voice source and emphasizes certain frequency regions known as formants. The first two formants are mostly associated with the produced vowels while the next three formants are identified with the timbrel characteristics of the voice. An important concept here is the singer's formant that is a cluster formed by 3rd, 4th and 5th formants around 3 kHz and useful for audience to hear the singing voice over a loud orchestra.

Definition of Phonation Modes

Sundberg [1] proposed that phonation modes can be considered as a third expressive resource for singing voice along with pitch and loudness. They are characterized by varying levels of glottal adduction that can affect the glottal flow resistance which can be estimated as the ratio of subglottal pressure and transglottal airflow [12, 15]. A high glottal adduction force results in a long closed phase and a low peak flow amplitude during the opening phase. An insufficient adduction force, on the other hand, can cause the vocal folds not closing completely; therefore, no closing phase can be observed but the obtained peak flow amplitude is high. Changing the level of glottal adduction within the phonatory adductory range can be used when choosing a phonation mode [15].

Sundberg [12] categorizes phonation modes into four groups 2: breathy, flow, neutral and pressed. Here, we introduce these four modes, and describe their characteristics in terms of glottal adduction force.

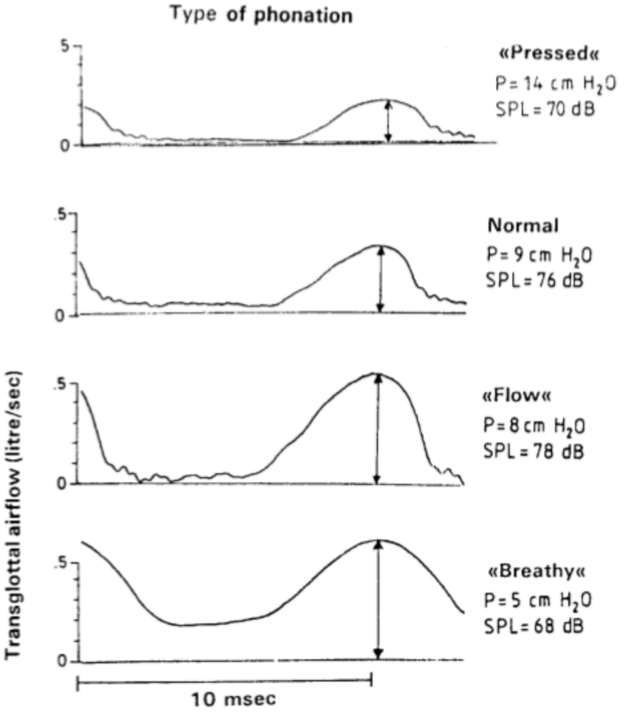

Figure 2: Examples of glottogram for different modes of phonation. On the right of the curves are given subglottal pressure (P) and sound pressure level (SPL) at 0.5 m distance. (Figure taken from [16], used with the permission of the author)

Pressed phonation is observed when a high glottal adduction force is used [1]. It is a hyperfunctional type of phonation that corresponds to a high subglottal pressure and a low transglottal airflow. It is typically associated with an elevated larynx and is considered to be poor in vocal efficiency due to not gaining much in the sound level but using a strong adduction force [17]. Pressed voice is also considered to be potentially harmful in phonotraumatic injuries because of the amount of the adduction force [18].

Neutral phonation can be achieved by using a lower glottal adduction force that results in a lower subglottal pressure and a higher transglottal airflow. The vocal efficiency is higher than in pressed phonation, and its characteristics resemble of speech [19].

Flow phonation is identified by using the lowest level of glottal adduction force that produces vocal fold closure [15]. Compared to neutral phonation, a lower subglottal pressure and a higher transglottal airflow is observed. The vocal efficiency is the highest in flow phonation as the ratio of the gained sound level and the amount of effort is higher compared to other phonation modes [17]. It is typically produced with a lowered larynx [17] and used as the standard phonation mode in Western Classical singing due to its effectiveness on allowing various resonances [19].

Breathy phonation results from an insufficient glottal adduction force that causes the vocal folds not to be closed completely [1]. It is a hypofunctional type of phonation that is characterized by a low subglottal pressure and a high transglottal airflow. When using breathy phonation, a high level of noise due to excessive transglottal airflow can be observed [20]. Breathiness is considered as an important factor of various pathological conditions [20] but some degree of incomplete glottal closure can be regarded as a normal laryngeal configuration [21].

In order to illustrate the differences among phonation modes in performances, Proutskova et al. [22] provide examples of uses of phonation modes. Pressed phonation can be heard in the forceful voice of James Brown in “I Feel Good”. An example for flow phonation can be the resonant voice of Liza Minelli in “New York, New York”. Breathy phonation is used to reflect sweetness and sexuality, and the performance of Marilyn Monroe in “Happy Birthday Mr. President” can be given as a famous example.

An important point to make is that the aforementioned phonation modes describe specific vocal fold closure and opening patterns when using the modal register. Due to the fundamental differences in sound production with the falsetto register, these descriptions of phonation modes may differ.

Related Work

For analyzing and identifying phonation modes, several research have been conducted. Grillo and Verdolini [18] used physical measurements to demonstrate that the ratio of subglottal pressure and transglottal airflow can be used to separate pressed, neutral and breathy modes. Millgård et al. [23] reported that closing quotient of the glottis and the difference between amplitudes of the first two harmonics in voice source spectrum are correlated with the amount of phonatory pressedness.

In order to have information about phonation modes on an audio signal processing perspective, a number of features have been designed and studied. Alku et al. [24] proposed that Normalized Amplitude Quotient (NAQ) feature that indicates the glottal closing phase can be used for distinguishing breathy, pressed and neutral modes in speaking. Sundberg et al. [25] reported that 73% of the variations in perceived pressedness can be explained by the variations of NAQ. Hillenbrand et al. [20] introduced Cepstral Peak Prominence (CPP), a feature that shows correlation with perceived breathiness. Harmonics-to-Noise Ratio (HNR), Jitter and Shimmer features were studied by Wakasa et al. [26] to analyze their usefulness for separating pressed and neutral modes.

Automatic classification approaches for phonation modes in singing have emerged with the curation of the first publicly available dataset [27] designed for this purpose. Proutskova et al.[22] argued against the use of spectral features such as Mel Frequency Cepstral Coefficients (MFCC) for this task and proposed a system using inverse filtering method to estimate the glottal source waveform in order to perform automatic classification. The proposed system uses inverse filtering derived features, such as NAQ, and a Support Vector Machine (SVM) model for the machine learning part. The resulting accuracy scores are in the range of 60% to 75%. A second approach, by Ioannidis et al. [28], is using Linear Predictive Coding (LPC) related features, and demonstrates a mean F-measure of 0.841 on the same dataset. Their method uses amplitudes of harmonics, formants, their differences and CPP as features, and a Logistic Model Tree algorithm as the classifier. Stoller and Dixon [29] extended the feature space used in the first 2 automatic classification attempts with others such as MFCC, and performed an analysis on the behavior of features against phonation modes. They reported that a simple rule-based method using Temporal Flatness, 0th coefficient of MFCC and CPP features resulted in 78% accuracy. A second classification attempt by Stoller and Dixon considered various feature sets and a feed-forward neural network, and achieved a mean F-measure of 0.868. The results of their analysis showed that the spectral features, e.g. MFCC, may have more relevant information than some other features from previous works, e.g. NAQ.

In order to extend the scope of phonation modes research, Rouas and Ioannidis [30] have curated a second dataset, in the same fashion as the first, containing recordings of a male baritone singer. Their work compares performance of two feature sets they describe as “Acoustic Descriptors” and “Glottal Features”. The first set includes LPC related features as well as CPP and HNR while the second set includes features such as NAQ, Peakslope and Maximum Dispersion Quotient (MDQ). They reported that the performance of Acoustic Descriptors were higher than Glottal Features with using a K-star classifier. Combining both feature sets improved the performance of automatic classification, and on a dataset containing recordings from both the first and the second datasets, they reported a mean accuracy of 79%.

Goals and Structure of the Thesis

In this study, we aim to contribute to phonation modes research in two ways: improving the methods of previous works in order to achieve a higher classification performance and developing a visual feedback system for singing students and teachers to use.

In order to achieve a higher classification performance, we consider extending the feature space used in previous works to investigate further information about the characteristics of phonation modes. The analyses of the features we propose to use present relevant information for the task. Moreover, with using essential machine learning steps such as feature selection and hyperparameter tuning with cross validation, we target higher performance scores obtained from automatic classification.

A visual feedback system can facilitate the understanding of phonation modes among singers. Our goal on developing such system is to take the first step toward a general feedback system on phonation modes that can be used by singing students and teachers in various institutions.

Contributions

The principle contributions of this thesis are as follows:

- Two new curated datasets designed for automatic classification purposes in the context of phonation modes in singing
- The feature space used in previous works is extended with PLP, RASTA-PLP and LFCC features, and based on the analysis results for this new set of features, their usefulness for the automatic classification of phonation modes in singing task is proved
- The first open source visual feedback prototype for learning/teaching phonation modes in singing is designed, and it is evaluated by singers and non-singers
- A conference paper presented in 15th Sound and Music Computing Conference: Yesiler, F. & Ramirez, R. A Machine Learning Approach to Classification of Phonation Modes in Singing. 15th Sound and Music Computing Conference (SMC 2018), (Limassol, 2018)

Chapter 2

Materials and Methods

In this chapter, we present the materials, e.g. datasets, toolboxes, and the proposed method used in our experiments. We start with an introduction to the datasets, and continue with the information regarding toolboxes we utilize. After, the steps outlining our proposed method, and the decisions on each individual step of the experiments are explained.

Datasets

In this section, we introduce the datasets used in our experiments. All the datasets contain recordings of single sustained vowels for various pitch ranges and phonation modes. Table 2.1 contains metadata information for the recordings in all the datasets.

For automatic classification purposes, we create balanced versions of each dataset in terms of the number of recordings for each phonation mode, and unless stated otherwise, we use only the balanced versions in our experiments.

Dataset-1

The first publicly available dataset for phonation modes in singing research (DS-1) is published by Proutskova et al. [22] The recordings are sung by a

Table 1: Metadata of datasets used in our experiments

	Vowels	Pitch Range			
		Breathy	Neutral	Pressed	Flow
DS-1	/a/, /e/, /i/, /o/, /ö/, /u/, /ü/, /ı/, /ä/	A3-G5	A3-G5	A3-C5	A3-G4
DS-2	/a/, /e/, /i/, /o/, /u/	C#2-G#4	D#2-D4	D#2-D4	D#2-E4
DS-3	/a/, /e/, /i/, /o/, /u/	A3-B5	A3-D#4 / G4-B5	A3-B5	A3-D#4 / G4-C6
DS-4	/a/, /e/, /i/, /o/, /u/	F#3-F4	F#3-F4	F#3-F4	F#3-F4

professional soprano singer. The pitch range of the recordings are between A3-G5 but not all the phonation modes are recorded in the entire pitch range. The recordings include 9 vowels that are used in various languages, and the total number of samples is 909.

In 2016, Proutskova announced that flow mode recordings do not represent the definition of Sundberg, and they should not be used for phonation modes related tasks[1].

Dataset-2

The second publicly available dataset for automatic classification purposes (DS-2) is published by Rouas and Ioannidis [30]. It contains recordings sung by a professional male baritone singer. There are 487 recordings, and the pitch range is C#2-G#4. For this dataset, 5 vowels that are used in Greek language (the singer's native language) are considered.

Dataset-3

To extend the scope of automatic classification of phonation modes, the first dataset we have curated (DS-3) contains recordings sung by a professional female soprano singer. 5 vowels that are used in Spanish language are recorded, and the pitch range is A3-C6. The recordings took place in Universitat Pompeu Fabra Poblenou Campus. Sennheiser 441 microphone is used, and there are

[1] https://osf.io/pa3ha/wiki/home/

515 recordings in total.

Dataset-4

The second dataset we have curated for this study (DS-4) contains recordings sung by a classically trained female soprano singer. As DS-3, the samples in DS-4 are recorded in Universitat Pompeu Fabra Poblenou Campus. In order to observe possible effects of recording conditions, a different microphone, AKG 4000, is used. The pitch range of DS-4 is F#3-F4, and 5 vowels that are used in DS-2 and DS-3 are considered. There are 240 recordings in total.

For both DS-3 and DS-4, before the recording sessions, the singers were briefed about phonation modes, and we conducted listening sessions where they heard the recordings from DS-1 and DS-2 as well as a number of popular examples given in previous works. The labeling of the recordings was done by the authors and the singers. Moreover, a computational analysis based on previous datasets is done to certify the correct use of phonation modes (e.g. if a feature shows higher values for breathy mode than neutral mode in DS-1 and DS-2, we search for the same patterns in DS-3 and DS-4).

Combined Dataset

The experiments in this study are performed on all datasets, separately and also combined (DS-C). Due to Proutskova's announcement regarding DS-1, we consider only DS-2, DS-3 and DS-4 for DS-C.

Toolboxes

In this section, we introduce the toolboxes that are used for feature extraction, feature selection and automatic classification steps of our experiments. The detailed information of the aforementioned steps is presented in Section 2.3.

RASTAMAT

RASTAMAT toolbox [31] is created by Dan Ellis, and it includes feature extraction algorithms and helper functions written in MATLAB. The algorithms in RASTAMAT are mainly used in speech-related tasks, and the contents of this toolbox are free.

MIRToolbox

MIRToolbox [32] contains a large number of algorithms written in MATLAB for facilitating computational approaches in Music Information Retrieval field. It follows an object-oriented design approach and is a free toolbox under GNU General Public License.

Praat

Praat [33] is a free toolbox designed to provide computational analysis of speech and phonetics. It provides various functions for tasks like formant analysis, speech synthesis etc. It is written in C and C++, and is published under GNU General Public License.

ProsodyPro

ProsodyPro [34] is a Praat script to be used in speech-related tasks. It facilitates the extraction of various features including formant and harmonics information, energies of different spectral bands etc.

Weka

Weka [35] is a toolbox containing a wide range of machine learning algorithms. It includes tools for data preprocessing, classification and visualization. It is published under GNU General Public License and written in Java.

Proposed Method

Our proposed method follows a supervised classification approach which is commonly used in many Music Information Retrieval (MIR) tasks. Figure 3 presents the steps of our approach.

Feature Extraction

In previous works on automatic classification of phonation modes in singing, various features such as amplitude differences of harmonics, CPP and MFCC are analyzed

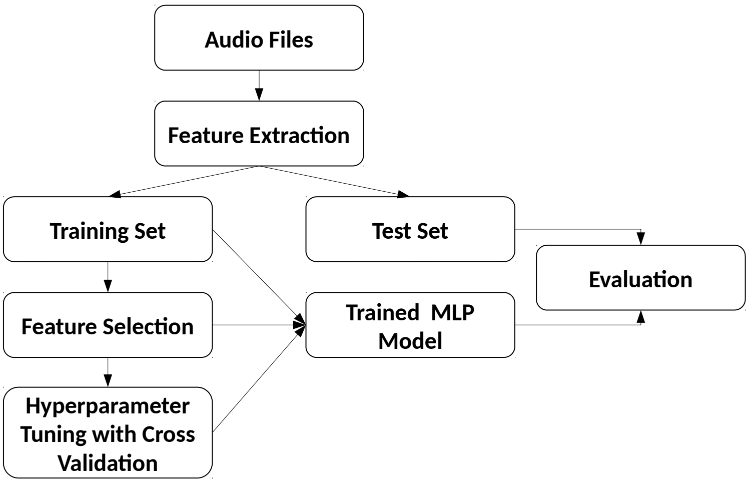

Figure 3: Main steps of the proposed method

in order to detect correlations between them and phonation modes, and they are used for supervised classification approaches. Our proposed method extends the feature space used in previous works and utilizes a feature selection algorithm to obtain a selected subset of features. Table 2 presents the entire list of features used in this study.

Table 2: Feature space of the proposed method

	Name	No of Features	Works Used		Name	No of Features	Works Used
1	MFCC & ΔMFCC	52	[29]	**9**	CPP	1	[28],[29],[30]
2	PLP & ΔPLP	52	-	**10**	Center of Gravity	1	[29]
3	RASTA-PLP & ΔRASTA-PLP	52	-	**11**	Formant Dispersion 1-3	1	-
4	LFCC	32	-	**12**	Hammarberg Index	1	-
5	Temporal and Spectral Flatness	2	[29]	**13**	Jitter & Shimmer	2	-
6	Spectral Flux	2	[29]	**14**	Harmonicity	1	[28],[29],[30]
7	Formants 1-3 Information	10	[28],[29],[30]	**15**	Energy Below 500Hz & 1000Hz	2	-
8	H1-H2, H1*-H2*, H1-A1, H1-A3	4	[22],[28],[30]	**16**	Energy Profiles of 500 Hz Wide Bands	15	-

1-Mel-Frequency Cepstral Coefficients (MFCC)

Extracting MFCC is one of the most commonly used techniques for retrieving spectral information. They are calculated by taking a Fourier transform of a windowed frame, mapping the powers of the obtained spectrum onto the mel scale with triangular bands, taking logs of the powers for the mel frequencies and using discrete cosine transform on the obtained log powers.

2-Perceptual Linear Prediction (PLP)

PLP is a spectral feature that uses various concepts from psychophysics of hearing. PLP coefficients are calculated by performing critical band analysis (bark scale warping) on the power spectrum, applying an Equal Loudness pre-emphasis and an amplitude compression by using Intensity-Loudness Power Law, and using Inverse Discrete Fourier Transform (IDFT) to be used by an all-pole model using the autocorrelation method.

3-Relative Spectral Transform - Perceptual Linear Prediction (RASTA–PLP)

RASTA-PLP is computed similarly to PLP but introduces a special band-pass filter for each frequency sub-band after the bark scale warping.

4-Linear Frequency Cepstral Coefficients (LFCC)

LFCC are computed very similarly to MFCC, and the only difference is instead of using a mel scale for triangular bands, LFCC algorithm uses equal spaced triangular bands. This approach facilitates capturing relevant cepstral information in the high frequencies [36].

5-Temporal and Spectral Flatness

Temporal and Spectral Flatness are the ratios between the geometric mean and the arithmetic mean of the signal in time and frequency domains, respectively.

6-Spectral Flux

Spectral Flux is computed as the euclidean distance between the power spectrum of a frame and the previous frame.

7-Frequency, Intensity and Bandwidth of 1st, 2nd and 3rd Formants (A1-I1-B1, A2-I2-B2, A3-I3-B3)

These features represent the amplitudes, intensities and bandwidths in the frequency domain of the first 3 formants. We also include a descriptor (No. of Formants) that is the mean of the number of formants found in each frame by Praat toolbox.

8-Amplitude Differences of Harmonics and Formants (H1-H2, H1*-H2*, H1-A1, H1-A3)

H1-H2, H1-A1 and H1-A3 is calculated as the amplitude difference between 1st and 2nd harmonics, 1st harmonic and 1st formant, and 1st harmonic and 3rd formant, respectively. H1*-H2* is the amplitude difference between first two harmonics formant adjusted [37].

9-Cepstral Peak Prominence (CPP)

CPP is a feature designed to detect breathiness in voice [20]. The notion behind CPP is that a highly periodic signal presents a well defined harmonic structure; thus a more prominent cepstral peak. To measure the "prominence", a linear regression line is computed between 1 millisecond and the maximum quefrency and the difference between the cepstral peak and the corresponding value on the regression line for the quefrency of the cepstral peak is computed.

10-Center of Gravity

Center of Gravity is calculated in the spectral domain.

11-Formants Dispersion 1-3

This descriptor is the average distance between adjacent formants up to F3.

12-Hammarberg Index

Hammarberg Index is calculated as the difference in maximum energy between 0kHz-2kHz and 2kHz-5kHz.

13-Jitter and Shimmer

Jitter and Shimmer are defined as mean absolute cycle-to-cycle difference divided by mean period and mean amplitude, respectively.

14-Harmonicity

Harmonicity is a measure of acoustic periodicity. It is based on the ratio of harmonically related energy to noise in audio signals.

15-Energy Below 500Hz and 1000Hz

These features represent the energy of voiced segments below 500Hz and 1000Hz.

16-Energy Profiles of 500Hz Wide Bands

15 energy profiles are computed from overlapping, 500Hz wide frequency bands, i.e. 0Hz-500Hz, 250Hz-750Hz, . . . , 3500Hz-4000Hz.

In addition to the features used in previous works, we propose to extend the scope of phonation modes research with features such as PLP, RASTA-PLP and LFCC. The reason for extending the feature space is to get as much information as possible from the audio. The first automatic classification study by Proutskova et al. [22] argued against using spectral features such as MFCC; however, Stoller and Dixon [29] reported that spectral features may have relevant information regarding this task. In our study, with computational analyses of features, we aim to explore possible correlations between other spectral features and characteristics of phonation modes.

The reason why we consider PLP, RASTA-PLP and LFCC features in specific is that they are compared with MFCC in a number of MIR tasks such as speech recognition [38] and speaker recognition [36]. The works comparing those features report that for various cases, the performance of those features can be higher than the others; thus, there is no “best feature” among them. Furthermore, the reason why we consider the features 11-13, 15, 16 is that they are included in the output of ProsodyPro script; therefore, the feature values are already calculated as a result of using the script.

For frame based features, e.g. MFCC, PLP and so on, the mean and standard deviation values are computed. For representation purposes, the rule for the abbreviations we use for the names of features in the next sections is summarized in Table 3.

Table 3: Abbreviations used for feature names

Feature	Δ (if used)	Coefficient	Mean / Standard Deviation	Examples
MFCC (M) **PLP (P)** **RASTA-PLP (R)** **LFCC (L)**	D	# of Coefficient	M or S	M1M PD1S RD10S L9M

In order to extract the aforementioned features, we use the toolboxes presented in Section 2.2. Table 4 presents the information regarding the choice of toolboxes for extracting each feature. In most cases, we use the default parameter values for each algorithm but for reproducibility purposes, the non-default parameter values that we use are shared in Appendix A.

Table 4: Choice of toolboxes for feature extraction

	Features
RASTAMAT	1-3
MIRToolbox	5,6
Praat	4,7
ProsodyPro	8-16

Stoller and Dixon [29] stated that for MFCC features, using the entire recordings including the voice onsets and releases results in a better performance while for the other features, they used trimmed versions of the recordings. In our proposed method, we use the entire recordings for features 1-3, 5 and 7, and for features 4, 6, 8-16, only the middle 600 milliseconds of the recordings are used. In order to facilitate trimming and labeling, some modifications to the source codes of the algorithms are made; however, this changes do not affect the algorithms we use for feature extraction.

Feature Selection

In order to follow a supervised learning automatic classification approach, one of the main problems is to decide on a feature subset representative for the task. With various features, one can extract a large amount of information about audio samples, but choosing the most relevant ones is an essential step of using machine learning techniques. Throughout the years, many feature selection algorithms have been designed to increase the performance of training the machine learning models, and depending on the task, the choice of using which algorithm may change.

Our proposed method utilizes Correlation-based Feature Selection (CFS) algorithm designed by Hall [39]. The main idea behind CFS algorithm is that the selected features should be highly correlated with classes while uncorrelated with each other. This approach presents a computationally efficient and scalable solution for discarding the redundant features. CFS algorithm is implemented as a part of Weka toolbox.

Automatic Classification

To perform automatic classification, our proposed method uses a Multi-Layer Perceptron (MLP) model. MLP is a type of artificial neural networks that utilizes a supervised learning method called backpropagation. It is composed of an input layer, a number of hidden layers and an output layer to make a prediction about the input. Activation functions of the hidden layers and the output layer may be different. Our decision on using an MLP model for our experiments is based on the approach of Stoller and Dixon [29].

In our experiments, we first divide the dataset into two parts in a stratified way regarding the classes: a training subset (90% of the instances) and a test subset (10% of the instances). The feature selection and hyperparameter tuning for the MLP model are performed on the training subset without any knowledge on the test subset in order to discard any potential bias.

On the training subset, CFS algorithm is applied to select the most relevant feature subset. With using the instances of the training subset and the feature vector selected by CFS algorithm, we perform 10-fold cross validation for tuning the hyperparameters of the MLP model. 10-fold cross validation divides the training subset into 10 stratified subsets, and for 10 iterations, assigns one of the subsets as the validation set, and evaluates the model that is trained with the other subsets on the validation set. Performance measures such as F-measure and accuracy score are calculated based on those 10 iterations. For the hyperparameter values, we take 0.01 as learning rate, 0.5 as momentum coefficient and 2000 as the number of epochs for the MLP model. To select the best performing hidden layer size, we use one hidden layer with (9, 10, 11, 12) nodes for individual datasets and (9, 10, 11, 12 ,13 ,14, 15) nodes for the combined dataset.

After performing 10-fold cross validation using different hidden layer sizes, the best performing value based on F-measure score is selected to be used for creating the MLP model. The entire training subset is used for training, and the obtained model is evaluated on the test subset. Since the instances in the test subset were kept apart for the feature selection and cross validation steps, the experiment is expected to be without any bias.

We perform 10 iterations of the aforementioned steps for each dataset. Selection of hidden layer size with cross validation step is performed only for the first iteration. The selected hidden layer size from Iteration 1 is used for the next 9 iterations. After getting F-measure and accuracy score values for each iteration, we calculate the mean values for each performance measure.

Chapter 3

Results

Feature Analyses

Before presenting the results of automatic classification with the extracted features, here, we discuss the performance of RASTA-PLP, PLP and LFCC features on distinguishing the phonation modes. For the analyses, we first conduct an Analysis of Variance (ANOVA) test on DS-C, and we determine the 10 features that have the highest F-scores. With the obtained 10 features, we perform post-hoc tests in order to observe the usefulness of those features on separating each phonation mode pair, e.g. Breathy-Neutral, Breathy-Pressed. For the post-hoc tests, we use Tukey's Honest Significant Difference (Tukey's HSD) test. The results of the analyses show that various RASTA-PLP, PLP and LFCC features present relevant information for the task; thus, our decision on extending the feature space is justified.

RASTA-PLP

Table 5 presents the results of our analysis for a selected subset of RASTA-PLP coefficients. The second and third columns show the F-score and p-values obtained from ANOVA, and the rest of the columns contain p-values obtained from Tukey's HSD test for each pair of phonation modes.

The analysis results present p-values lower than 0.05 for ANOVA and Tukey's HSD

Table 5: ANOVA and Tukey's HSD results for selected RASTA-PLP features

	F-score	p-value	B-N	B-P	B-F	N-P	N-F	P-F
RD1M	95.91	<0.01	<0.01	<0.01	<0.01	0.04	<0.01	<0.01
R2M	79.00	<0.01	<0.01	<0.01	<0.01	0.59	<0.01	<0.01
R2S	74.20	<0.01	<0.01	<0.01	<0.01	0.04	0.06	0.73
R0M	69.91	<0.01	<0.01	0.04	<0.01	0.04	<0.01	<0.01
RD3M	65.46	<0.01	<0.01	<0.01	<0.01	0.34	<0.01	<0.01
R1M	59.36	<0.01	<0.01	<0.01	<0.01	0.04	<0.01	>0.9
RD3S	58.58	<0.01	<0.01	<0.01	<0.01	>0.9	<0.01	0.01
RD1S	54.53	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	<0.01
R5M	44.15	<0.01	<0.01	<0.01	0.04	<0.01	<0.01	0.23
RD2M	42.76	<0.01	<0.01	>0.9	<0.01	<0.01	<0.01	<0.01

tests for RD1M, R0M and RD1S; therefore, the distribution of these features for all phonation modes are significantly different from each other. This shows that these features can be used for separating all the phonation modes from each other.

For distinguishing breathy mode, all the presented features except RD2M can be used based on their distributions; however, as for separating neutral mode from the others, only 6 features, RD1M, R0M, R1M, RD1S, R5M and RD2M, present relevant information. In terms of identifying the flow mode recordings, a different subset of 6 features, RD1M, R2M, R0M, RD3M, RD1S and RD2M, can be used. For the pressed mode, only aforementioned 3 features that can be used to separate all phonation modes from each other present useful information.

In Figure 4, we present the values of the selected subset of RASTA-PLP features for every recording in DS-C, separated by their phonation modes and sorted by their pitch values. For representation purposes, the values are normalized with L2 norm within each feature. Since DS-C is a combined dataset consists of recordings from DS-2, DS-3 and DS-4, deviations within each phonation mode caused by having different datasets can be observed as vertical lines.

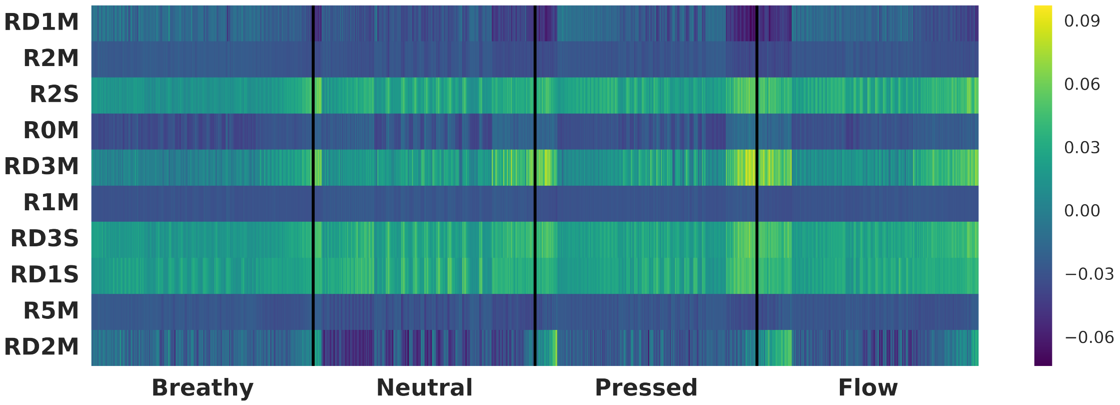

Figure 4: Values of a selected subset of RASTA-PLP features from DS-C

PLP

Table 6 presents the analysis results of a selected subset of PLP features, and is organized in the same fashion as Table 5. For the 10 PLP features that have the highest F-scores, p-values obtained from ANOVA are lower than 0.01. For P2S, the p-values obtained from Tukey's HSD test are lower than 0.05 for each phonation mode pair; thus, the distributions of these features for each phonation mode are significantly different than the others.

Table 6: ANOVA and Tukey's HSD results for selected PLP features

	F-score	**p-value**	**B-N**	**B-P**	**B-F**	**N-P**	**N-F**	**P-F**
P2S	98.33	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.03
PD1S	91.97	<0.01	<0.01	<0.01	<0.01	0.26	<0.01	<0.01
PD3S	64.58	<0.01	<0.01	<0.01	<0.01	>0.9	0.02	<0.01
PD2S	62.85	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.80
PD2M	50.16	<0.01	<0.01	<0.01	>0.9	<0.01	<0.01	<0.01
PD6M	40.11	<0.01	<0.01	<0.01	0.11	<0.01	<0.01	<0.01
P2M	34.34	<0.01	<0.01	<0.01	<0.01	<0.01	0.58	<0.01
PD6S	32.96	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.59
P0M	32.85	<0.01	<0.01	<0.01	0.03	<0.01	>0.9	<0.01
P8M	22.01	<0.01	<0.01	<0.01	<0.01	0.58	>0.9	0.87

For separating breathy mode from the others, all the features in the selected subset with exceptions of PD2M and PD6M can be used. Distributions of PD2M and PD6M show that these features are useful for identifying neutral and pressed modes. While

PD2S and PD6S can also be used for distinguishing neutral mode from the others, P0M can be used for distinguishing pressed mode. Lastly, for identifying flow mode, PD1S and PD3S present relevant information along with P2S.

Figure 5 illustrates the values of the aforementioned PLP features for each recording and is organized in the same fashion as Figure 4.

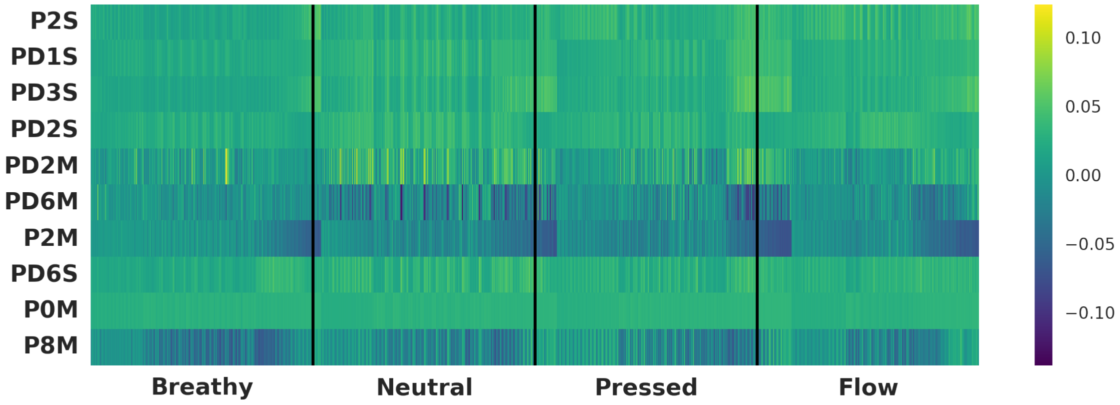

Figure 5: Values of a selected subset of PLP features from DS-C

LFCC

The results of our analysis on LFCC features can be seen in Table 7. All the highest performing 10 features have p-values from ANOVA test lower than 0.01; however, for all the mode pairs, only L1M has p-values lower than 0.05.

Table 7: ANOVA and Tukey's HSD results for selected LFCC features

	F-score	p-value	B-N	B-P	B-F	N-P	N-F	P-F
L1M	106.94	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
L7M	52.51	<0.01	<0.01	<0.01	<0.01	0.02	>0.9	0.06
L10M	39.83	<0.01	<0.01	<0.01	0.05	<0.01	0.12	<0.01
L2M	39.82	<0.01	<0.01	<0.01	<0.01	0.04	0.30	<0.01
L9M	34.11	<0.01	<0.01	<0.01	<0.01	0.07	>0.9	0.02
L8M	31.85	<0.01	<0.01	<0.01	<0.01	0.15	0.08	>0.9
L4S	20.17	<0.01	<0.01	<0.01	<0.01	>0.9	<0.01	0.02
L11M	13.18	<0.01	0.11	<0.01	0.16	<0.01	>0.9	<0.01
L2S	11.91	<0.01	<0.01	<0.01	<0.01	>0.9	0.88	0.88
L16M	11.21	<0.01	0.51	<0.01	<0.01	<0.01	0.09	0.38

Along with L1M, 6 features, L7M, L2M, L9M, L8M, L4S and L2S, can be used for separating breathy mode from the others. L2M, L10M and L11M are useful for identifying pressed mode, and L4S presents relevant information for distinguishing flow mode from the others.

Figure 6, organized in the same way as Figure 4 and 5, shows the values of selected LFCC features for each recording in DS-C.

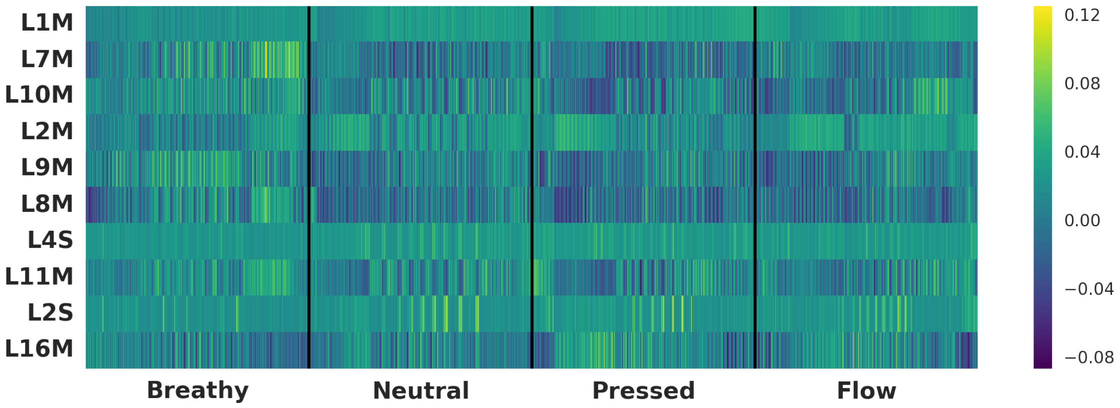

Figure 6: Values of a selected subset of LFCC features from DS-C

Selected Features from Previous Work

DS-1 is the only dataset that was used in the works of Proutskova et al. [22], Ioannidis et al. [28], and Stoller and Dixon [29] to analyze the behavior of features and to perform automatic classification, and some of the results they achieved can be specific for a particular singer, gender or voice type. Rouas and Ioannidis [30] evaluate the performance of Acoustic and Glottal features on both DS-1 and DS-2; however, they did not perform an analysis on the distributions of those features for any of the datasets. Figure 7 illustrates distributions (max, min and quartiles) of a selected set of features used in previous works, scaled between 0 and 1. For separating pressed mode, M1M and H1-H2 demonstrate relevant information on DS-1 while the same does not apply on DS-2. CPP is useful for identifying breathy mode on both datasets but the relationships of distributions of neutral and pressed modes differ. HNR presents lower results for breathy mode on DS-1; however, the same correlation is not observed on DS-2. Therefore, this analysis can be seen as a justification of our decision on creating new datasets for feature analyses and automatic classification tasks.

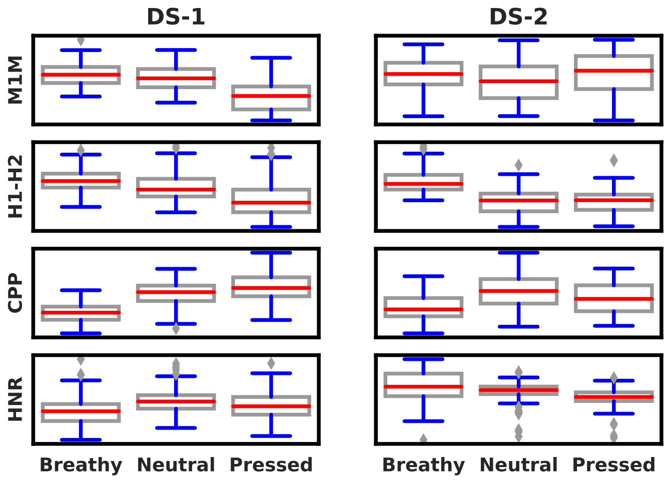

Figure 7: Distributions of selected set of features on DS-1 and DS-2

Selected Features from CFS Algorithm

After the feature extraction step, training and test subsets are created for 10 iterations from each individual and the combined dataset, and CFS algorithm is used to select the best performing feature subset for each training subset. Due to the random nature of creating training subsets, the obtained feature subsets differ from the ones in other iterations. In this subsection, we report our findings on the selected features for the training subsets obtained from DS-C.

The average number of selected features from CFS algorithm is 40.8 with the maximum number being 48 and the minimum number 35. To be able to have an understanding of the output of CFS algorithm, here, we present the features that are selected by CFS algorithm for all of the 10 iterations. Table 8 presents the common selected features with their F-scores and p-values obtained from ANOVA for each individual feature. We observe that a number of features used for this task for the first time (e.g. PLP, RASTA-PLP, LFCC) are included in the output of CFS algorithm for all the iterations; thus, the results are consistent with our feature analyses in the previous subsections.

Table 8: ANOVA results of the feature subset selected by CFS algorithm in every iteration

Features	F-score	p-value	Features	F-score	p-value
Spectral Flatness	113.04	<0.01	**RD2M**	42.76	<0.01
EP 2.5k-3k	112.49	<0.01	**L2M**	39.82	<0.01
L1M	106.94	<0.01	**L9M**	34.11	<0.01
P2S	98.33	<0.01	**MD8S**	30.71	<0.01
RD1M	95.91	<0.01	**EP 1k-1.5k**	23.81	<0.01
EP 2.25k-2.75k	96.63	<0.01	**MD9S**	21.65	<0.01
Hammarberg Idx	94.07	<0.01	**PD1M**	18.62	<0.01
CPP	84.23	<0.01	**P5M**	17.47	<0.01
MD5S	65.77	<0.01	**MD6M**	14.95	<0.01
R1M	59.36	<0.01	**M0M**	13.88	<0.01
L7M	52.51	<0.01	**H1-H2**	12.32	<0.01
M4S	49.11	<0.01	**Jitter**	5.65	<0.01
M5M	46.47	<0.01			

An important point for interpreting the usefulness of the presented feature subset is that in each iteration, there are other features included in the output of CFS algorithm; therefore, the features in Table 8 are not sufficient to achieve the F-measure and accuracy scores we present in Section 3.2. On the other hand, this feature subset proves its importance for the task since it has been selected for 10 different randomly created training subsets.

Automatic Classification

The obtained mean F-measures and the mean accuracy scores for all the datasets are presented in Table 9 and Table 10, respectively, in comparison to previous works. It can be observed that our decisions on extending the feature space and using a feature selection algorithm considerably increased the performance of methods proposed in previous works. The standard deviation of the F-measure values for DS-1, DS-2, DS-3, DS-4 and DS-C are 0.036, 0.021, 0.031, 0.097 and 0.030, respectively, and the standard deviation of the accuracy scores for the same datasets are 0.036, 0.021, 0.031, 0.100 and 0.030, respectively.

In Table 11, we present the resulting aggregated confusion matrix for 10 iterations on

Table 9: F-measure values of the proposed method in comparison to previous works

Dataset	DS-1	DS-2	DS-3	DS-4	DS-C
Proutskova et al.[22]	-	-	-	-	-
Ioannidis and Rouas[28]	0.841	-	-	-	-
Stoller and Dixon[29]	0.868	-	-	-	-
Rouas and Ioannidis[30]	-	-	-	-	-
Proposed Method	0.897	0.972	0.922	0.855	0.903

Table 10: Accuracy score values of the proposed method in comparison to previous works

Dataset	DS-1	DS-2	DS-3	DS-4	DS-C
Proutskova et al.[22]	60-75%	-	-	-	-
Ioannidis and Rouas[28]	-	-	-	-	-
Stoller and Dixon[29]	-	-	-	-	-
Rouas and Ioannidis[30]	81.62%	88.51%	-	-	-
Proposed Method	89.81%	97.21%	92.29%	85.83%	90.26%

DS-C. The most confused mode pair is Neutral-Pressed, and 15 and 14 recordings of flow mode are misclassified as neutral and pressed, respectively. The highest accuracy is observed with breathy mode.

Table 11: Aggregated confusion matrix obtained from DS-C. Rows and columns represent true and predicted classes, respectively

	Breathy	Neutral	Pressed	Flow
Breathy	276	3	1	7
Neutral	7	249	23	7
Pressed	2	24	256	6
Flow	6	15	14	257

Chapter 4

PhonationRT: A Visual Feedback Prototype

The main goal of this chapter is to introduce our visual feedback software prototype, PhonationRT, for detecting phonation modes in singing to be used for various applications. PhonationRT is written in C++, and it uses several free audio libraries and toolboxes for feature extraction and automatic classification. The prototype is licensed under Affero General Public License v3 (Affero GPLv3).

The variations in the mode of phonation in singing can be used as an expressive resource in performance practices [1]. Phonation modes are characterized with varying levels of glottal adduction force, and they can be controlled independently from pitch and loudness. As a result of using different levels of glottal adduction force, Sundberg [12] proposes 4 categories of phonation modes: breathy, flow, neutral and pressed.

The principle goal of this prototype is to be used in music education. Methods and applications for tracking pitch, timing and loudness are well designed and commonly used by singing students and/or teachers. Students can practice by themselves with those applications to see whether they are hitting the right note at the right time or not, and this may facilitate the learning process of students in most cases. An important point regarding these assessment/feedback systems is that they can be used as reinforcement tools for education, and it is not advised to learn how to sing only with those applications. Learning and teaching how to sing have a pedagogical aspect, and singing teachers can prevent their students developing unhealthy singing habits. Therefore, the system we propose should not be considered as a standalone tool for learning how to sing the right phonation modes but it should be a helper tool for students and teachers.

System Architecture

The main components of PhonationRT can be seen in Figure 8. In this section, we describe the materials and methods used in each step of our prototype.

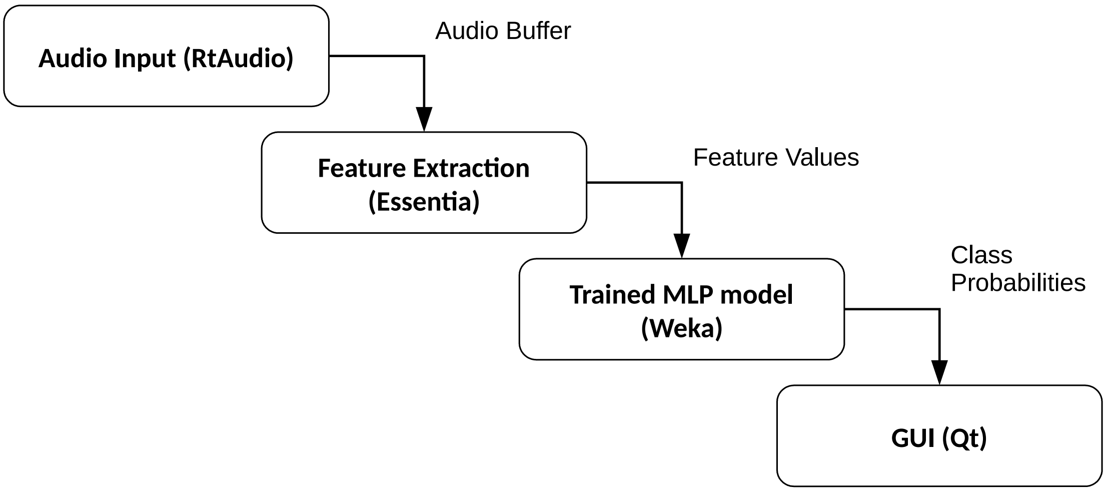

Figure 8: Main components of PhonationRT prototype

Audio Input

PhonationRT prototype utilizes RtAudio [40], a set of C++ classes that facilitates real time audio input/output by providing a common API across Linux, Macintosh and Windows operating systems. It is developed and being maintained by Gary P. Scavone and other developers, and employs an object oriented design. RtAudio is published under copyright that allows the rights to use, copy, modify, merge etc. under certain conditions. We use RtAudio to obtain real time audio input for our system.

Feature Extraction

In our prototype, feature extraction step is being performed by algorithms included in Essentia [41], an open source C++ library, developed by Music Technology Group (MTG) at Universitat Pompeu Fabra, for audio analysis and audio-based MIR tasks. It contains a vast collection of algorithms that implement standard digital signal processing blocks, a large set of audio descriptors etc. Essentia supports Linux, Macintosh and Windows operating systems as well as iOS and Android, and it is released under Affero GPLv3.

To the obtained audio input, we first apply YIN algorithm in order to estimate pitch and pitch confidence. Our prototype does not extract other features and give visual feedback unless the pitch confidence value is above 0.7.

For developing a prototype that works in real time, we cannot extract all the features we use in our proposed method; therefore, we select some of the features to use in PhonationRT based on their computational costs. When a pitch is detected, we use Windowing, Spectrum, MFCC, Flatness and Centroid algorithms of Essentia to calculate 12 MFCC coefficients (1-12, excluding 0th coefficient) as well as Temporal Flatness, Spectral Flatness and Spectral Centroid coefficients. We use a circular buffer of 10 frames to calculate mean values for all the coefficients and standard deviation values for only MFCC coefficients.

Automatic Classification

For the machine learning step of our prototype, we use an MLP model created by Weka. The model is trained with DS-C, and for the non-default hyperparameters, we take 0.01 as learning rate, 0.5 as momentum coefficient, 2000 as number of epochs and 1 hidden layer with 17 nodes. The F-measure obtained by 10-fold cross validation is 0.864. The model is hard-coded into our prototype, and based on the values of the input vector, it predicts the probabilities of that input being a member of one of the four phonation modes.

An important point to mention regarding automatic classification is that Weka uses a normalization step for the feature values to scale them between 1 and -1. We use the same normalization method with respect to the feature values in DS-C.

Visual Feedback

After obtaining the probabilities for each phonation mode, we use Qt [1], a software development framework, to provide a visual feedback for the user. Qt framework is being developed both by The Qt Company and the Qt Project, and works on Linux, Macintosh and Windows as well as Android and iOS. It is dual-licensed under commercial and open source (LGPLv3, GPLv2 and GPLv3) licenses .

An example of the feedback can be seen in Figure 9. In this case, we see the visual feedback of an audio frame that is predicted to be an example of a breathy mode with 0.05 probability, flow with 0.85, neutral with 0.00 and pressed with 0.10.

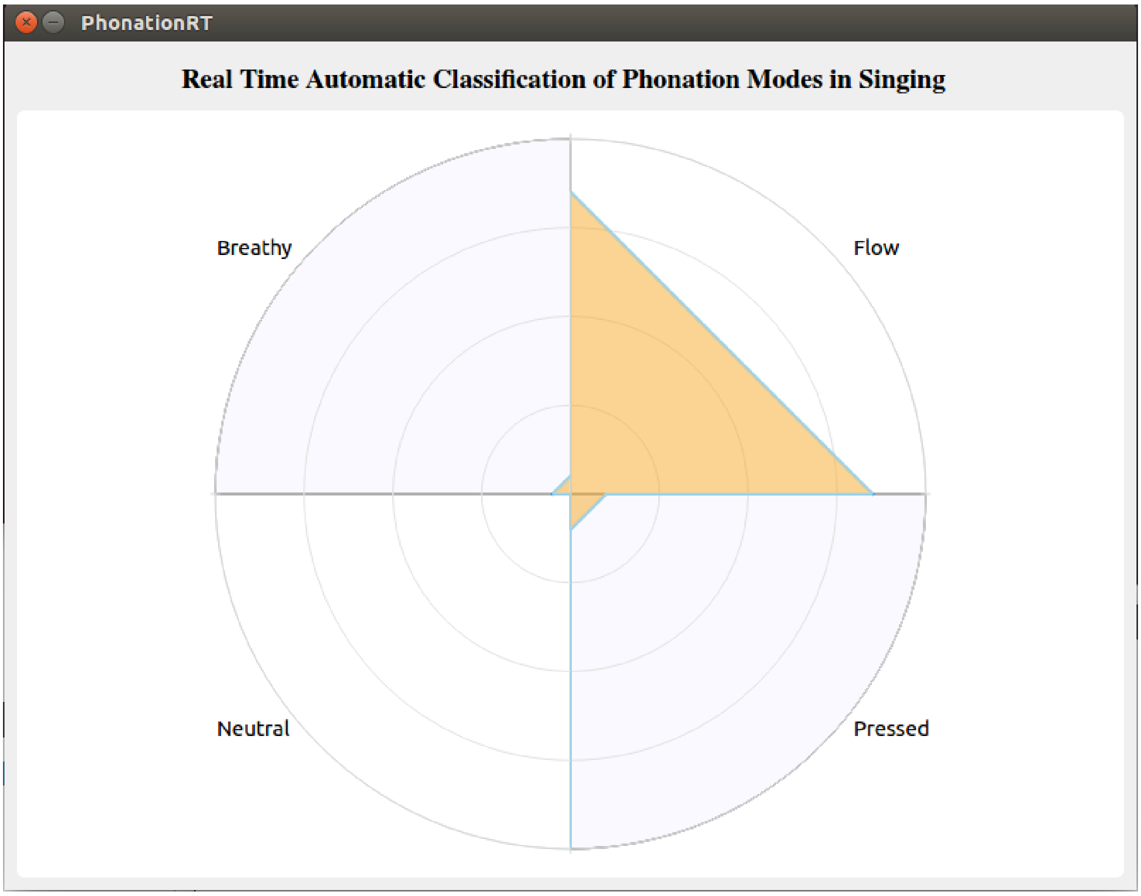
PhonationRT
Real Time Automatic Classification of Phonation Modes in Singing
Breathy
Flow
Neutral
Pressed

Figure 9: An example of the visual feedback obtained from PhonationRT prototype

[1] https://www.qt.io/

Prototype Evaluation

For the evaluation of our prototype, we have conducted individual user studies with singing students/teachers and non-singers.

In each session, we explained the phonation modes and played some examples from our datasets and popular recordings. After a short briefing on the concept of phonation modes, we asked the participants to test our prototype by singing short notes and phrases with different phonation modes. The tests are followed by a short questionnaire for the participants to evaluate the system. The list of questions can be seen in Table 12. We asked the participants to give ratings between 1 and 5, 1 being the lowest and 5 being the highest. We also included an optional commentary part for each question for the participants to elaborate their answers.

Table 12: Questions asked to participants for PhonationRT evaluation

	Questions
Q1	How close is the feedback to your sound perception?
Q2	How informative do you find the feedback?
Q3	How useful do you find the idea of such a feedback system for self learning?
Q4	How useful do you find PhonationRT prototype for self learning?
Q5	How useful do you find the idea of such a feedback system for teaching?
Q6	How useful do you find PhonationRT prototype for teaching?

In total, there were 5 participants (1 singing teacher, 2 singing students and 2 non-singers) for the prototype evaluation sessions. Age of the participants vary from 23 to 59, and we had 3 male and 2 female participants. The average music listening hours per week of singers is 19.3 while it is 3.5 for non-singers. The number of years of singing training for singers is are 3, 4 and 42, and the average number of singing hours is 3.3. The genres/styles of singing are Western Classical, Turkish Folk and Carnatic.

The average ratings for each question can be seen in Table 13. The user comments for the first question indicate that detecting flow mode is the least accurate, and one user stated that the results appear to be pitch and vowel dependant. For the second question, two of the users commented that instead of giving feedback for each frame, using a linear feedback that has the x-axis representing time and y-axis representing phonation mode can be more informative.

Table 13: Average ratings given by participants

	Average ratings
Q1	2.8
Q2	2.8
Q3	4
Q4	2.8
Q5	3.8
Q6	2.6

Based on the ratings and comments for questions 3 and 5, we can reach the conclusion that the idea of a visual feedback system for learning/teaching phonation modes can be useful. An important point is that as our participants stated, such feedback system should not be used as a standalone tool for learning but it can be complementary to an education with a singing teacher.

Lastly, the user responses for questions 4 and 6 show that the prototype is not useful for learning/teaching as is. These answers are in line with the responses for questions 1 and 2. User comments indicate that future improvements in the accuracy of the feedback and the informativeness of the interface would lead to a more useful system to be used for educational purposes. A discussion on the current weaknesses of the system can be seen in Chapter 5.

Chapter 5

Discussion and Conclusion

In the last chapter of this thesis, our primary goal is to give a brief summary of our work with a discussion regarding our proposed method and the obtained results. Moreover, we define 5 challenges for automatic classification of phonation modes in singing research, and propose them as possible future work for this field. Lastly, we give a short description of the "Open Science" concept and include a link to an online repository containing the necessary information and documents to reproduce our experiments.

Summary

Datasets

In order to extend the scope of phonation mode research, we have recorded two new datasets that contain single-sustained vowel recordings of two female soprano singers for this work, and labeling of the obtained recordings are done by the authors and the singers. Supervised learning approaches for MIR tasks require labeled data, and Flexer and Grill [42] emphasize the importance of Inter-rater Agreement while preparing ground truth annotations for any kind of data. For the labeling of the new datasets, a further study that consists of many different annotators may be conducted in order to improve the reliability of the ground truth annotations.

In terms of recording conditions, we have used the same recording room for both recording sessions but we chose different types of microphones for each dataset. Our decision on this is based on creating a variation on the recording conditions so that any potential bias resulting from using the exact same type of equipment would be partially discarded. Moreover, since one of our goals is to develop a visual feedback prototype for educational purposes, we are aware that the equipments the users have may not match for all cases; therefore, minor variations in recording conditions may increase compatibility issues of the system.

Another point regarding the use of the datasets in our experiments is that as mentioned in Section 2.1, we created balanced versions of each dataset in order to have equal amounts of recordings for each phonation mode. Since we leave out some of the recordings for our experiments, different decisions on which recordings to include in the balanced datasets may have an effect on the performance scores obtained from automatic classification step of our method.

Feature Extraction

One of our goals in this study is to expand the feature space used in previous works for automatic classification of phonation modes. Based on the work of Stoller and Dixon [29] that favors the usefulness of MFCC for this task, we have considered other spectral features such as PLP, RASTA-PLP and LFCC for expanding the feature space. These features are compared with MFCC for various MIR tasks such as speech recognition [38] in previous works.

An important point regarding feature extraction step is that in most cases, we use the default values for the parameters of the feature extraction algorithms that are included in the toolboxes we use. A parameter optimization step for the algorithms may be performed to increase the performance of automatic classification but our proposed method does not employ such measures. Furthermore, there exist different implementations for the same features in different toolboxes, and our approach does not compare such alternatives.

Feature Analyses

RASTA-PLP, PLP and LFCC

In order to justify our decision on expanding the feature space used in previous works, we performed analyses of the new features we propose to use for this task. Based on ANOVA and Tukey's HSD test, we see that a number of new features present useful information for phonation modes research. Here, we provide a short summary of our analyses for the 10 features within each group, e.g. RASTA-PLP, PLP, LFCC, that have the highest F-scores from ANOVA.

Based on Tukey's HSD test, for distinguishing breathy mode from the others, 9 out of 10 RASTA-PLP features, 8 out of 10 PLP features and 6 out of 10 LFCC features can be used. For separating neutral mode, 6 RASTA-PLP features, 5 PLP features and 1 LFCC feature present useful information while for identifying pressed mode, 3 RASTA-PLP features, 5 PLP features and 4 LFCC features can be considered. Lastly, for differentiating flow mode, 7 RASTA-PLP features, 3 PLP features and 2 LFCC features can be used.

Features Used in Previous Works

The second part of our feature analyses is performed in order to justify our decision on creating new datasets. Having only one dataset to analyze the behavior of certain features may result in problematic conclusions that may not be obtained from other data. Our analysis points out that the behavior of some features may change due to other factors that may not be related to variations in phonation modes, and first 3 automatic classification attempts only used one dataset. In order to have a more general understanding of usefulness of features on phonation modes in singing, we suggest gathering more data in order to facilitate this line of research.

Features Selected by CFS algorithm

Machine learning algorithms are designed to find patterns in data and to make sense of the new data using those patterns. An important point to consider is the ratio of the number of attributes and the number of instances. A high number of attributes may cause the machine learning algorithm to overfit the training data; therefore, a decrease in the performance of the trained model on a new set of data, e.g. test set, may be observed. To avoid overfitting, it is often beneficial to use a feature selection algorithm, and in our proposed method, we consider using CFS algorithm because of its efficiency in terms of performance vs computational costs. Other feature selection algorithms may produce different performance scores in automatic classification step.

The average number of features selected from training subsets of DS-C by CFS algorithm is 40.8. Considering that the size of our feature space is 231, and the number of instances in training subsets obtained from DS-C is 1033, CFS algorithm can be considered useful to avoid the aforementioned overfitting problem.

Automatic Classification

The mean performance scores obtained from automatic classification step show that our proposed method results in considerably higher scores compared to previous works. We use an MLP model for the classification but many other algorithms such as Support Vector Machines can be used for this task. The scope of this thesis does not include comparisons of various machine learning algorithms, and the performance scores may increase/decrease as a result of employing different models.

The aggregated confusion matrix of 10 iterations shows that our proposed method tends to suffer a weakness for confusing neutral and pressed modes. To tackle this issue, features that are robust in segregating these modes may be more emphasized in future works.

PhonationRT

User evaluation studies for PhonationRT demonstrate that although the idea of such visual feedback system is useful for learning/teaching phonation modes in singing, the prototype is not useful enough to be used in educational purposes as is. We can discuss this issue in three main topics: the scope and the specifications of the training data, parameters used in the system architecture and the interface.

In terms of the scope of the training data, as mentioned earlier, we need to extend the training data with recordings from singers who have different voice types and specialize in different genres/styles of singing. Our training data contain recordings from classically trained singers; thus, examples of phonation modes used in other singing styles become hard to detect. Moreover, all the recordings for our training data are made in well-equipped recording studios with high-quality microphones. For the prototype evaluation sessions, the recording conditions of the training data were not reproduced, and, in our case, using a built-in laptop microphone certainly affects the resulting spectrum of the input signal. Using microphones that have better frequency responses may increase the user ratings on accuracy of the prototype.

Another issue is the software parameters (sampling rate, buffer size, circular buffer size) used in our prototype. The parameter values are chosen as a result of initial experiments done by the authors, and we do not perform any parameter tuning steps for users. We see this point as a potential improvement area for future works.

Lastly, based on the user comments, another area to improve may be the interface of the prototype. Our frame-based visual feedback was rated average in terms of informativeness, and possible improvements include presenting the history of the performance instead of only the current frame and performing an offline score analysis based on a pre-annotated singing piece. More advancements may be achieved with conducting user studies focused on User Experience / User Interaction Design aspects.

Challenges and Future Work

In this section, we list a number of the challenges regarding automatic classification of phonation modes, and we define an outline regarding future work.

- **Challenge 1:** Making use of various spectral features that are hard to interpret is becoming a more widely accepted approach in supervised learning based MIR tasks. An important challenge regarding this aspect is to be able to obtain sufficient amount of data in order to justify the behavior of features against variations in phonation modes.

- **Challenge 2:** As a part of singing voice research, the use of phonation modes in various musical styles and genres is not well studied to this day. By developing more accurate machine learning models, studies regarding commonly used phonation modes in certain styles/genres can be facilitated.

- **Challenge 3:** By looking at an ethnomusicological point of view, our knowledge of phonation modes obtained from computational analyses can be used to investigate cultural preferences for using a certain mode of phonation. As an example, Proutskova et al. [22] analyzed the correlations between the choice of phonation modes and the status of women in a society. Such studies can strengthen the connection between humanistic and scientific research.

- **Challenge 4:** In order to facilitate the adoption of the concept of phonation modes into singing education, a standalone real time feedback software can be designed. The essential considerations toward this goal are having a large amount of data to be able to generalize the conclusions and carefully optimizing the models that will be used in this software.

- **Challenge 5:** Along with pitch and loudness, the mode of phonation is considered as a third expressive resource of the singing voice; however, in scores using western musical notation, there are guidelines for only pitch, timing and loudness for practicing and performing songs. A notation system for phonation modes can be designed and included in western music notation to help the singers understand which phonation mode should be used in a certain part of a song.

Reproducibility

The Open Science movement is becoming more emphasized and getting widely accepted in the scientific community, and it has changed the ways of exchanging information. The main principle is that the gained scientific knowledge should not be kept in private and instead, it should be shared in order to facilitate advancements in science.

In order to support the "Science 2.0" revolution and facilitate future works regarding phonation modes research, all the resources we use in this thesis are shared. Toolboxes we use for our proposed method are chosen from freely available software. The explanation of the steps and our decisions throughout our experiments are documented and distributed in the following Github repository:

https://github.com/furkanyesiler/PhonationModes-MasterThesis

List of Figures

List of Tables

Bibliography

[1] Sundberg, J. What's so special about singers? *Journal of Voice* **4**, 107–119 (1990).

[2] Juslin, P. & Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* **129**, 770–814 (2003).

[3] Lerdahl, F. & Jackendoff, R. S. *A Generative Theory of Tonal Music* (MIT Press, 1985).

[4] Sundberg, J. The KTH synthesis of singing. *Advances in Cognitive Psychology* **2**, 131–143 (2006).

[5] Randel, D. M. & Apel, W. *The new Harvard dictionary of music* (Belknap Press, 1986).

[6] Simpson, J. A. & Weiner, E. *Oxford English Dictionary* (Clarendon Press, 1989), 2 edn.

[7] Palmer, C. & Hutchins, S. What is musical prosody? In Ross, B. (ed.) *Psychology of Learning and Motivation*, vol. 46, 245–278 (Academic Press, The address of the publisher, 2006).

[8] Sundberg, J., Iwarsson, J. & Hagegård, H. A singer's expression of emotions in sung performance. *STL-QPSR* **35**, 81–92 (1994).

[9] Juslin, P. N. & Sloboda, J. A. *Music and Emotion* (Oxford Univ. Press, 2008).

[10] Darwin, C. *The Expression of the Emotions in Man and Animals* (John Murray, 1872).

[11] Scherer, K. R. Expression of emotion in voice and music. *Journal of Voice* **9**, 235–248 (1995).

[12] Sundberg, J. *The Science of the Singing Voice* (Northern Illinois University Press, 1987).

[13] Sundberg, J. *Röstlära* (Konsultfirma Johan Sundberg, 2007).

[14] Gramming, P. Vocal loudness and frequency capabilities of the voice. *Journal of Voice* **5**, 144–157 (1991).

[15] Herbst, C. T., Hess, M., Müller, F., Švec, J. G. & Sundberg, J. Glottal adduction and subglottal pressure in singing. *Journal of Voice* **29**, 391–402 (2015).

[16] Sundberg, J. Vocal fold vibration patterns and phonatory modes. *STL-QPSR* **35**, 69–80 (1994).

[17] Sundberg, J. Vocal fold vibration patterns and modes of phonation. *Folia Phoniatrica et Logopaedica* **47**, 218–228 (1995).

[18] Grillo, E. U. & Verdolini, K. Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects. *Journal of Voice* **22**, 546–552 (2008).

[19] Thalén, M. & Sundberg, J. Describing different styles of singing: A comparison of a female singer's voice source in "Classical", "Pop", "Jazz" and "Blues". *Logopedics Phoniatrics Vocology* **26**, 82–93 (2001).

[20] Hillenbrand, J., Cleveland, R. A. & Erickson, R. L. Acoustic correlates of breathy vocal quality. *Journal of Speech Language and Hearing Research* **37**, 769–778 (1994).

[21] Murry, T., Xu, J. J. & Woodson, G. E. Glottal configuration associated with fundamental frequency and vocal register. *Journal of Voice* **12**, 44–49 (1998).

[22] Proutskova, P., Rhodes, C., Crawford, T. & Wiggins, G. Breathy, resonant, pressed automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research* **42**, 171–186 (2013).

[23] Millgård, M., Fors, T. & Sundberg, J. Flow glottogram characteristics and perceived degree of phonatory pressedness. *Journal of Voice* **30**, 287–292 (2016).

[24] Alku, P., Bäckström, T. & Vilkman, E. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America* **112**, 701–710 (2002).

[25] Sundberg, J., Thalén, M., Alku, P. & Vilkman, E. Estimating perceived phonatory pressedness in singing from flow glottograms. *Journal of Voice* **18**, 56–62 (2004).

[26] Wakasa, K., Matsubara, M., Hiraga, Y. & Terasawa, H. Acoustic characteristics of pressed and normal phonations in choir singing by male singers. In *Proc. of the 2017 Int. Symposium on Musical Acoustics*, 136–139 (Montreal, 2017).

[27] Proutskova, P., Rhodes, C., Crawford, T. & Wiggins, G. Breathy or resonant - A controlled and curated dataset for phonation mode detection in singing. In *Proc. of the 13th Int. Society for Music Information Retrieval Conf. (ISMIR 2012)*, 589–594 (Porto, 2012).

[28] Ioannidis, L., Rouas, J.-L. & Desainte-Catherine, M. Caractérisation et classification automatique des modes phonatoires en voix chantée. In *XXXèmes Journées d'études sur la parole* (Le Mans, France, 2014).

[29] Stoller, D. & Dixon, S. Analysis and classification of phonation modes in singing. In *Proc. of the 17th Int. Society for Music Information Retrieval Conf. (ISMIR 2016)*, 80–86 (New York City, 2016).

[30] Rouas, J.-L. & Ioannidis, L. Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings. In *Interspeech 2016*, 150–154 (San Francisco, 2016).

[31] Ellis, D. P. W. PLP and RASTA (and MFCC, and inversion) in Matlab (2005). URL `http://www.ee.columbia.edu/ln/rosa/matlab/rastamat/`. [Accessed 24- August- 2018].

[32] Lartillot, O. & Toiviainen, P. A Matlab toolbox for musical feature extraction from audio. In *Proc. of the 10th Int. Conference on Digital Audio Effects* (Bordeaux, 2007).

[33] Boersma, P. & Weenink, D. Praat: Doing phonetics by computer [Computer program] (2018). URL `http://www.praat.org/`. [Accessed 24- August- 2018].

[34] Xu, Y. ProsodyPro - A tool for large-scale systematic prosody analysis. In *Proc. of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10 (Aix-en-Provence, 2013).

[35] Hall, M. *et al.* The WEKA data mining software: an update. *SIGKDD Explorations* **11**, 10–18 (2009).

[36] Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C. & Shamma, S. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 559–564 (Waikoloa, 2011).

[37] Iseli, M., Shue, Y.-L. & Alwan, A. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America* **121**, 2283–2295 (2007).

[38] Këpuska, V. Z. & Elharati, H. A. Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model classifier in noisy conditions. *Journal of Computer and Communications* **3**, 1–9 (2015).

[39] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, The University of Waikato (1999).

[40] Scavone, G. P. RtAudio: A cross-platform C++ class for realtime audio input/output. In *Proc. of the 2002 Int. Computer Music Conference (ICMC 02)*, 196–199 (Goteborg, 2002).

[41] Bogdanov, D. *et al.* ESSENTIA: an audio analysis library for Music Information Retrieval. In *Proc. of the 14th Int. Society for Music Information Retrieval Conf. (ISMIR 2013)*, 493–498 (Curitiba, 2013).

[42] Flexer, A. & Grill, T. The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research* **45**, 239–251 (2016).

Appendix A

Non-default Parameter Values for Feature Extraction

Table 14: Non-default values for extracting MFCC with RASTAMAT

Parameters	**Values**
lifterexp	0
sumpower	0
preemph	0
maxfreq	8000
nbands	80
useenergy	1

Table 15: Non-default values for extracting LFCC with Praat

Praat function	**Parameters**	**Values**
To LPC (autocorrelation)	**Timestep**	0.01
To LFCC	**Number of coefficients**	16