Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings

Jean-Luc Rouas, Léonidas Ioannidis

Automatic classification of phonation modes in singing voice: towards singing style characterisation and application to ethnomusicological recordings

Jean-Luc Rouas[1], *Leonidas Ioannidis*[1]

[1]LaBRI - CNRS UMR 5800

rouas@labri.fr, lioannid@labri.fr

Abstract

This paper describes our work on automatic classification of phonation modes on singing voice. In the first part of the paper, we will briefly review the main characteristics of the different phonation modes. Then, we will describe the isolated vowels databases we used, with emphasis on a new database we recorded specifically for the purpose of this work. The next section will be dedicated to the description of the proposed set of parameters (acoustic and glottal) and the classification framework. The results obtained with only acoustic parameters are close to 80% of correct recognition, which seems sufficient for experimenting with continuous singing. Therefore, we set up two other experiments in order to see if the system may be of any practical use for singing voice characterisation. The first experiment aims at assessing if automatic detection of phonation modes may help classify singing into different styles. This experiment is carried out using a database of one singer singing the same song in 8 styles. The second experiment is carried out on field recordings from ethnomusicologists and concerns the distinction between "normal" singing and "laments" from a variety of countries.

Index Terms: singing analysis, phonation, voice quality

1. Introduction

We often categorize our music collection according to the singer, and often we recall musical pieces by their lyrics and the accompanying voice. Thus, the singing voice is what gives the music its identity by providing it the meaning that no other instrument can give. Among the ways to characterize the singing voice, one of the most salient features is the vocal quality. Indeed, vocal quality is roughly speaking the color of the voice. Since the lyrics and the partition usually have to be respected by the singer, his identity and the feelings he wants to add to his interpretation have to be expressed through modulations of the voice quality (as opposed to speech, where a speaker may also adapt his choice of words, their duration and the intonation patterns). In this study, we will focus more precisely on a particular expression of voice quality in singing voice, which is namely the phonation modes. After a brief explanation on the phonation modes, we will describe the databases we used and how we built a system to classify phonation modes on isolated vowels. Then, another experiment is carried out on continuous singing to assess if phonation modes may be linked to singing styles.

2. Phonation modes

This papers follows and extend the work of Proutskova [1, 2]. It is based on the assumption that there exists four main phonation modes in singing [3]. These phonation modes are namely: breathy, neutral (or modal), flow (or resonant) and pressed.

For example, the breathy voice may be used to express sexuality or sweetness, the most common example being Marilyn Monroe singing "happy birthday". Flow phonation may be encountered in very "active" singing as for example in the "belting" technique often used by Aretha Franklin. A good example for pressed phonation would be James Brown in "I feel good".

In a more technical understanding, phonation modes result primarily from the adjustments made at the larynx level. Both Laver [4] on speech and Sundberg [3] on singing voice define the phonation as having three dimensions: i) the pitch, ii) the loudness and iii) the laryngeal adjustments.

The main cues that can be used for describing the four phonation modes are:

- *Pressed* phonation is associated with an elevated larynx position which also affects the vocal tract shape. There is also a stronger muscular tension around the larynx. The pressed voice is very rich in harmonic content as it favors the rise of the harmonics rather than the fundamental. Pressed-ness in voice may be perceived as a fatigue as the phonation becomes ineffective and can, under some circumstances, be part of vocal health problems.
- In *Breathy* phonation there is a relaxing of the musculature that is responsible for the adduction/abduction of the vocal folds. There is a reduced vocal fold adduction and minimal vocal fold impact stress [3, 5]. The result is a laxed voice with a high level of noise from the turbulent air that passes through. Noise to Harmonic ratio is expected to be generally higher than in the other phonations and significantly higher on the spectral region above 2KHz [6]. Another strong perceptual indicator of breathyness is the sensation of excessive laryngeal airflow [7].
- *Flow* voice is defined more as a vocal technique as it is used exclusively in singing, and unlike the other modes it requires vocal training. Sundberg suggests that flow voice is typically produced by a lowered larynx [8], while in [9] it is proposed that flow voice results from the condition where the vocal tract impedance is considerably smaller than the glottal impedance in an effort to optimize the mean glottal resistance or in other words the vocal output power. The produced loudness seems to be the key issue in this vocal type where the goal is to achieve higher levels of loudness with the minimum effort. In this procedure there are three elements that characterize this phonation: 1) Formant tuning, especially the first, 2) ample harmonic content and 3) narrowing of the laryngeal vestibule [10]

- In *Modal* voice, also known as normal phonation, we find a full vibration of the vocal folds, along their entire length.

3. Isolated vowels datasets

3.1. The soprano dataset

We obtained a freely distributed database which was build specifically for the study of the phonation modes [1]. This database consists of 790 short recordings of nine different vowels, found in Russian language, the singers native language. The singer is a female soprano singer with musical training. The average duration of the recorded samples is 1,34 seconds, varying from 0,9 until 1,6 seconds long.

3.2. The baritone dataset

As this database consists only of recordings of one female soprano singer, we decided to extend it. We thus recorded a male baritone singer. As for the soprano dataset, this database is also made freely available upon request.

The recordings took place at SCRIME, Studio de Creation et de Recherche en Informatique et Musique Electroacoustique[1].

The singer, Georgios Papaefstratiou, is a professional trained baritone singer of the Choir of the National Opera of Bordeaux[2]. Besides his training in lyrical singing he also has experience in popular music and general music training. His singing vocal range is G2 - B4, but we recorded only the notes in the range of A2 - G4 which he claimed to be his usual working range. Before recording the database, the singer was briefed on the subject. A listening procedure followed where he heard the recordings of the soprano database in order to be familiarized with the goals and aims of the specific project of phonation mode detection.

The recordings were made using a Neumann U87 Ai studio microphone with all the filter and attenuation switches turned off. The cardioid function was chosen to attenuate reflections from the recording room. We tried to reproduce as mush as possible the experimental settings found in [1], however due to the specificity of the singer, the pitch range was shorter, from A2 to G4 for all phonation types and the singer recorded only five vowels /a/, /o/, /e/, /i/, /u/ found in his native language (greek). The database consists of 487 samples of an average duration of 1.43 seconds.

4. Conception of the automatic classification system on isolated vowel samples

Both databases (soprano and baritone) were downsampled to 16kHz. We decided to select only the 500ms middle section of the sung vowels in discard to avoid attack and release parts.

4.1. Feature extraction

A number of audio descriptors are extracted from the audio signal. We separate these descriptors into two sub-sets, acoustic and glottal descriptors. These two sets will be evaluated separately and jointly in their ability to classify the different phonation modes. The first set consists of features extracted from the acoustic signal as recorded from the microphones and the second set consists of descriptors calculated after a glottal-wave was estimated using glottal inverse-filtering.

4.1.1. Acoustic Descriptors

A hanning window function of 25ms length has been used for all the descriptors that are presented here. A hop-size of 5ms overlapped at 1/5th of the window has been implemented. Wherever there is use of an short-term FFT (STFT) function, the FFT size is of 512 samples long, approximately 32ms, and zero-padded. One single value for each sound sample is extracted that is calculated from the mean values of all the samples in the STFT procedure. The final acoustic feature set is of dimension 24.

The different families of acoustic descriptors are:

- Harmonics Amplitude: It has been reported from many authors [11, 12, 3] that the difference between the two first harmonics in a source signal is an important parameter that strongly relates to specific types of phonation. Further more the third harmonic can also help determine the phonation to a lesser extent. These are calculated for the first three harmonics (H1, H2, H4).
- Formant Frequencies, Bandwidth and Amplitude: The formant frequencies are important parameters in human speech and voice signals in general. There are often used for vowel estimation and their parameters are highly tied to the vocal tract. Formants amplitudes are computed for the first three formants (A1, A2, A3), frequencies and bandwidth are calculated on the first four formants (F1, F2, F3, F4 and B1, B2, B3, B4).
- Harmonics & Formant Amplitude Differences: these features can describe the spectral shape with respect to the fundamental frequency amplitude. Amplitude differences are computed for H1-H2, H2-H4, H1-A1, H1-A2, H1-A3).
- Cepstral Peak Prominence (CPP): This method has been developed by [13] in an effort to characterize breathy voice in vocal signals for pathological speakers suffering voice disorders. Cepstral Peak Prominence (CPP) is based on periodicity measures on the cepstrum of the voice signal.
- Harmonic-to-Noise Ratio: The spectral shape is a feature that can characterize and differentiate well enough the phonation types when measured in the source signal [11]. HNR are computed for 4 different frequency bands: 0-0.5kHz, 0.5-1.5kHz, 1.5-2.5kHz, 2.5-3.5kHz.

4.1.2. Glottal Features

We refer to this set of features as glottal because they are extracted from the glottal waveforms estimated using glottal inverse-filtering.

The method used estimates vocal tract linear prediction coefficients and the glottal volume velocity waveform from a speech signal using Iterative Adaptive Inverse Filtering (IAIF) method. Analysis is carried out on a GCIsynchronous (CGI:glottal closure instant) basis and waveforms are generated using overlap and add. The method is described in [14]. The method is synchronized to the glottal closure instants, thus a GCI detection is needed before applying the inverse-filtering method. This method is described in [15] and in recently published comparative review of GCI and GCO detection methods was found to be most accurate [16].

[1] scrime.labri.fr

[2] http://www.opera-bordeaux.com

The features extracted for the estimated glottal signal are:

- Normalized Amplitude Quotient: The normalized amplitude quotient was introduced in [17] and was presented as a time-based parametrization method for a more robust measurement of the closing phase, than the closing quotient (CQ).
- Quasi-Open Quotient (QOQ): The quasi-open-quotient is a variation of the open-quotient. The open-quotient measures the ratio of the time in which the glottis remains open during phonation. The QOQ is expected to have a big value in laxed and breathy voice types.
- H1H2: The difference between the fundamental frequency energy (H1) and its first harmonic (H2) is measured from the source signal estimated with the inverse-filtering method.
- Parabolic Spectral Parameter (PSP): Parabolic spectral parameter is a frequency domain feature develloped by [18], for the quantification of the glottal volume velocity waveform.
- Peakslope: This feature has been proposed as an effective one for lax-tensed voice discrimination [11].
- Maximum dispersion quotient (MDQ): This parameter was proposed in [19] for the differentiation of breathy and pressed (tense) vocal types.

4.2. Automatic classification

We carried out experiments with different families of classifiers using the weka software [20] ans libsvm [21]. The results described below are given only for the Kstar classifier which was the best performing one.

4.3. Evaluation and Results

We first used the two datasets separately, then we merged the two datasets reaching a total of 1135 instances to classify. For all the evaluations a non-overlapping 10-fold cross-validation scheme was held for the training-testing procedure. This means that the whole dataset is divided randomly into 10 equally-sized subsets. At each iteration 9 sets are used for the training and the last set is used for testing. Finally, the results from the 10 iterations are collected to computed the overall performance.

Table 1: Summary of the accuracy results for the different feature sets. Results are in % of correct classifications.

	Acoustic	Glottal	Acoustic+Glottal
Soprano	79.74	64.90	81.62
Baritone	89.96	57.31	88.51
All	77.96	59.17	78.59

When trying to compare the two feature sets for their ability to well characterise the four phonation modes we can see that the acoustic set gives the best performance. Adding the glottal set to the acoustic one does not bring a significant change to the overall classification rate (significance test carried out using T-test).

The authors of the soprano database proposed in their work [2] a vowel-dependent scheme for the classification of the phonation. As a comparison, they reported an average accuracy of 65%, with the results varying from 52 to 75% according to the vowel. Our results are slightly better, with an overall classification rate of 81% correct.

The performances we obtained using only the baritone dataset are better than with the soprano dataset, the main errors for the soprano dataset are for quite high pitched samples (above C4) which may be linked to the fact that using different phonation modes can become difficult when the pitch increases.

Table 2: Confusion matrix for the classification experiment over the whole dataset and the acoustic features set. Results are expressed in %

	Breathy	Modal	Flow	Pressed
Breathy	88.49	10.61	0.29	0.58
Modal	12.02	74.57	6.87	6.52
Flow	0	4.16	75.75	20.07
Pressed	0.81	2.85	26.53	69.79

In table 2, the confusion matrix is given for the acoustic set only. We observe that breathy and modal phonation are well identified but sometime confused. Flow and pressed voices are distinguished from the other phonation modes but there is some confusion between them. In [2] is also reported that the main confusions are between breathy + modal and flow + pressed, a finding that was reproduced in our experiments. This experiment shows that a system can be designed to classify phonation modes using only acoustic information. In the next section, an experiment is carried out to assess whether this system may be useful for singing style characterisation.

5. Towards singing style classification

The goal behind this experiment is to observe if we could classify singing styles according to the detection of phonation modes. As training corpus we use the combined dataset of soprano and baritone (whole database used for training). Only the acoustic features set is used. In order to form our test dataset we use the singing styles dataset. The dataset is annotated according to the singing style by the singer itself and these classes are used as the ground truth.

5.1. Singing styles database

The singing style dataset consists of 32 recordings presented in [22]. This database consists of recording of a unique female singer whom performs parts of the song “Amazing grace”, the popular christian hymn written by the poet John Newton. The two first verses are sung several times in various singing styles. These styles are namely: RnB, Belting technique, Classical, Country, Jazz, Legato, Pop, Rock. The total duration of the database is 1103 seconds with an average of 39 seconds per file.

5.2. Extraction of the vocalic regions

In order to extract the vocalic regions which are the only regions where we can take a decision on the phonation modes, we used a basic speech alignment system. This system is built using HTK [23] and is trained on speech from the VoxForge project[3]. After the alignment, we only kept the vowels of length above 0.5 seconds.

5.3. Results and Discussion

The results for the classification task are displayed in table 3. We see that according to the singing style there are different

[3] voxforge.org

Style	Breathy	Modal	Flow	Pressed
R&B	2.90	**42.81**	3.55	**50.72**
Belting	1.07	1.95	**63.40**	33.56
Rock	0	3.28	**76.85**	21.79
Jazz	7.54	**78.22**	2.82	11.40
Pop	**18.21**	**50.29**	**12.43**	**19.06**
Legato	4.02	**73.65**	19.09	3.21
Country	15.55	**58.41**	0	**26.03**
Classical	5.52	**93.51**	0	0.96

Table 3: Classification of phonation mode according to singing style - results are expressed in % of total time

preferred phonations that have been detected. We can first remark that breathy voice although being the phonation that was better recognized in the above experiments has very low rates in all singing styles, probably because of its ornamental use rather than a singing style that is used constantly by the singers.

For RnB style we find that the singer chooses modal and pressed phonation equally. RnB music is characterized by high vocal effort so pressed phonation is not surprisingly high in this case. Belting and rock style shows a high rate of flow phonation which can be thought reasonable since flow and belting voice can be considered related voicing techniques. In jazz performances we observe a high rate of modal voice and a relatively high breathy rate, when compared to the rest of the singing styles. Pop performances are the ones where the biggest variety in the terms of phonation modes is observed, with modal being the most detected one. In legato and classical styles modal voice is chosen more often while in country style although modal phonation has the most instances there is a significant percentage of pressed and breathy.

6. Preliminary experiments on ethnomusicological recordings

The Diadems project (Description, Indexing, Accessibility to ethnomusicological and audio documents) is a project funded by the French National Research Agendy (ANR) and it aims at designing tools for the analysis and indexing of musical content adapted to the needs of ethnomusicological researchers. The main difficulties of such a database are that it consists of field recordings meaning that many things may happen in background (conversations, music, natural noises, etc.) and the recordings were made from the 1900 to nowadays using diverse hardware.

One of the challenges of the project is to study if some distinctions can be made automatically between different "levels" of singing (such as lament, chanting or singing) and different speaking styles (storytelling, recitation, talking). As a preliminary study, we decided to use the system to classify excerpts from the lament and singing classes. Lament is defined by the presence of several of the four common icons of crying (the cry break, the voice inhalation, the creaky voice and the falsetto vowel) proposed by [24]. These two categories of singing were found to be quite difficult to discriminate while they could be separated from the other classes [25].

Even if usual acoustic parameters do not help to make the distinctions between these two categories, we assume that since laments include extreme cases of phonations, there would be a lot more instances of breathy and pressed phonations, while in "normal" singing the distribution between the four phonation modes should be more balanced.

In that experiment, we considered 10 singing segments and 6 lament segments from different countries (Albania, Turkey, Paraguay, Ethopia, Azerbadjan, Vietnam, Lebanon, Mexico). The singing excerpts lasted for a total of 14 minutes and the lament excerpts 7 minutes. The proportions of the different phonation modes found are represented in figure 1.

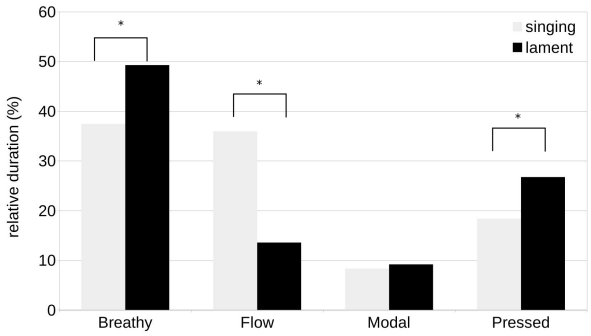

Figure 1: Proportions of phonation modes in the DIADEMS excerpts (in % of duration). * indicates significant differences

The results seems to confirm our hypothesis since the number of "extreme" cases of phonation is much more important in lament than in singing, the proportion of breathy and pressed phonation accounting to more than 75% of the total time. Differences between the distribution of phonation modes are highly significant ($p<0.01$) except for the modal phonation were there is no significant difference between the lament and singing voice.

However, this result is to be tempered by the fact that some singing styles may also involve a good proportion of breathy and pressed phonation. Considering the classification of individual recordings may thus prove ineffective using only these cues and require further research.

7. Discussion

In this paper, we have described a method that successfully automatically classified phonation modes using only parameters from the audio signal. This classification has been carried out on isolated vowels datasets, including a new baritone dataset, and achieved about 80% of correct answers. Therefore, we applied this system on "real" continuous singing data in order to assess its usefulness for two tasks: singing style characterisation and discrimination between singing and lament on ethnomusicological recordings. Although preliminary, the results obtained on these two tasks are encouraging us to further investigate in these directions. For the singing style characterisation, we need to apply the method to a greater number of singers but we have to find voice only recordings (no instrumental background) in different styles. For the ethnomusicological experiment, we wish to test the combination of the detection of phonation modes with other acoustic features.

8. Acknowledgements

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the "Investments for the future" Programme IdEx Bordeaux - CPU (ANR-10-IDEX-03-02) and the DIADEMS project (ANR-12-CORD-0022).

9. References

[1] P. Proutskova, C. Rhodes, and T. Crawford, "Breathy or Resonant a Controlled and Curated Database for Phonation Mode Detecion in Singing," *International Symposium on Music Information Retrieval*, no. Ismir, pp. 589–594, 2012.

[2] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, Resonant, Pressed - Automatic Detection Of Phonation Mode From Audio Recordings of Singing," *Journal of New Music Research*, pp. 1–28, 2012.

[3] J. Sundberg, *The science of the singing voice*. Northern Illinois University Press, 1987.

[4] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics, 1980.

[5] M. Garnier, N. Henrich, M. Castellengo, D. Sotiropoulos, and D. Dubois, "Characterisation of Voice Quality in Western Lyrical Singing: from Teachers' Judgements to Acoustic Descriptions," *Journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 62–91, Nov. 2007.

[6] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[7] E. U. Grillo and K. Verdolini, "Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects," *Journal of Voice*, vol. 22, no. 5, pp. 546–552, 2008.

[8] J. Sundberg, "Vocal fold vibration patterns and phonatory modes," *Folia Phoniatrica Logopedica*, vol. 47, pp. 218–228, 1995.

[9] I. R. Titze, "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 367–376, 2002.

[10] C. G. Smith, E. M. Finnegan, and M. P. Karnell, "Resonant voice: spectral and nasendoscopic analysis." *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 607–22, Dec. 2005.

[11] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform." in *INTERSPEECH*, 2011, pp. 177–180.

[12] P. Alku, "Glottal inverse filtering analysis of human voice production ? A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. October, pp. 623–650, 2011.

[13] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic Correlates of Breathy Vocal Quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[14] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[15] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals." in *Interspeech*, 2009, pp. 2891–2894.

[16] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 994–1006, 2012.

[17] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *the Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[18] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter - a new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[19] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, June 2013.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] C.-C. Chang and C.-J. Lin, "Libsvm : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[22] N. Henrich and L. Popeil, "Acoustical description of 8 common singing styles produced by a single female singer: preliminary results," in *Care of the Professional Voice Symposium*, Philadelphia, June 2003.

[23] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[24] G. Urban, "Ritual wailing in amerindian brazil," *American Anthropologist*, vol. 90, no. 2, pp. 385–400, 1988.

[25] L. Feugre, B. Doval, and M.-F. Mifune, "Using pitch features for the characterization of intermediate vocal productions," in *Proc of 5th International Workshop on Folk Music Analysis (FMA),*, 2015.