BREATHY OR RESONANT – A CONTROLLED AND CURATED DATASET FOR PHONATION MODE DETECTION IN SINGING

Polina Proutskova
Goldsmiths, University of London
proutskova@googlemail.com

Christophe Rhodes
Goldsmiths, University of London
c.rhodes@gold.ac.uk

Geraint Wiggins
Queen Mary, University of London
geraint.wiggins@eecs.qmul.ac.uk

Tim Crawford
Goldsmiths, University of London
t.crawford@gold.ac.uk

ABSTRACT

This paper presents a new reference dataset of sustained, sung vowels with attached labels indicating the phonation mode. The dataset is intended for training computational models for automated phonation mode detection.

Four phonation modes are distinguished by Johan Sundberg [15]: breathy, neutral, flow (or resonant) and pressed. The presented dataset consists of ca. 700 recordings of nine vowels from several languages, sung at various pitches in various phonation modes. The recorded sounds were produced by one female singer under controlled conditions, following recommendations by voice acoustics researchers.

While datasets on phonation modes in speech exist, such resources for singing are not available. Our dataset closes this gap and offers researchers in various disciplines a reference and a training set. It will be made available online under Creative Commons license. Also, the format of the dataset is extensible. Further content additions and future support for the dataset are planned.

1. MOTIVATION: NARROW, WIDE, BREATHY, RESONANT SINGING IN VARIOUS DISCIPLINES

Phonation modes play an important role in singing: they are an essential characteristic of a singing style – all musical traditions have cultural preferences for the use of one or more phonation modes; they are used as a means for expressive performance; they can be indicative of voice disorders; subtle changes in phonation mode production are used routinely by singing teachers to determine the progress of a student.

Johan Sundberg in his seminal work "The Science Of The Singing Voice" identifies four different phonation modes in singing: breathy, neutral, flow (called resonant by other authors) and pressed [15]. To illustrate the differences between phonation modes let us bring some well-known examples.

Breathy vocalisation is used skillfully by jazz and popular music singers to express qualities like sweetness or sexuality: think of Marilyn Monroe's most famous performances like "I wanna be loved by you"[1] or "Happy birthday Mr President"[2]; or listen to Ella Fitzgerald's and Louis Armstrong's "Dream a little dream of me"[3]. This mode of vocal production can easily be distinguished by human listeners from the flow phonation mode, such as a resonant, vibrating vocalising by Liza Minelli on her "New York, New York"[4]; and from the pressed phonation, e.g. the tense, forceful voice of James Brown in "I feel good"[5].

At the same time, Liza Minelli can be quite forceful; Ella Fitzgerald's vocals are very dominant but light and economical most of the time, because flow phonation is natural for her voice. All the singers mentioned above use flow phonation, and all of them have their personal preferences for varying phonation and using these variations as stylistic markers or as their individual expressive devices.

In Muslim countries of the Old High Culture a call for prayer can be heard five times a day from the minarets of the mosques. Call for prayer is an art form and Muezzins are experts in using squeezed, narrow sound (pressed phonation) in their singing which makes their performances very expressive. Their vocal technique is very different from a rounded, flying performance of a Western European Gregorian chant (neutral phonation) and is at the same time distant from the brassy, resonant Greek or Russian liturgic singing (flow phonation).[6]

[1] http://www.youtube.com/watch?v=MLU0jndUGg4
[2] http://www.youtube.com/watch?v=k4SLSlSmW74
[3] http://www.youtube.com/watch?v=j6TmogXhOZ8
[4] http://www.youtube.com/watch?v=rgusCINe260
[5] http://www.youtube.com/watch?v=XgDrJ5Z2rKw
[6] Here, again, situation with employed phonation modes can be quite ambiguous. For example, Greek Byzantine singers use somewhat pressed, nasalized phonation on higher pitches quite a lot. In fact, a Byzantine singer with

While the term phonation mode is borrowed from voice acoustics, the differentiation between breathy and pressed voices, between tense and open singing is operational in many voice-related research areas: ethnomusicology, singing education, medical research (phoniatrics, vocology) as well as in linguistics (phonetics). This is how an ethnomusicologist Alan Lomax describes the difference between narrow and wide vocalisation:

"The measure concerns the contrast between the voices which sound mellow, relaxed and richly resonant (we call this *wide*) and the voices which sound tense, pinched and restricted in resonance (which we call *narrow*). Many singing styles can be characterized as having one or the other; in some rare cases both may occur; and many ways of vocalizing (like everyday American speech) are neutral in width – these we call *mid*, singers with a "speech" tone." [10, p. 125]. Lomax then gives a number of examples: narrow singing from Indonesia and Thailand; wide, open singing from Eastern Europe; the mid mode form the US and from Ireland.

These examples demonstrate that breathy, pressed or resonant singing production can be representative of a singing style or even a music culture. While each voice is different and two singers never sing the same way, every musical tradition displays cultural preferences for the use of particular phonation mode(s), which are imposed on the singers performing in this tradition. In many cases a single phonation mode is encouraged: for example baritone singers in Western classical music are trained to sing in flow phonation and move through their singing career using just this phonation mode. In contrast, in classical Ottoman tradition a singer was expected to operate in all four phonation modes.

Apart from being a cultural characteristic, breathy or tense vocalisation can be indicative of vocal disorders: hypofunction and hyperfunction of the glottis [5]. Their diagnostics and treatment are a prime concern in the disciplines of vocology (voice habilitation) and phoniatrics (in case of functional or anatomic pathologies) [13].

While in some examples even a less experienced listener can easily distinguish between various phonation modes, in other cases this distinction can be very subtle and requires training and expertise to be identified correctly. Lomax refers to his narrow vs wide singing descriptor (he calls this descriptor vocal width or vocal tension) as an "emotionally loaded quality" and thus explains why some people have difficulties in rating it [10]. Vocal width is one of 36 descriptors of the Cantometrics system – a global parametrisation of world's singing styles. Lomax and his Cantometrics team manually rated more than 5000 recordings of singing from around the world. Of all 36 descriptors vocal width appeared to be the hardest to rate consistently: the inter-rater consensus scores for this descriptor are the lowest (see [10], p. 168). Victor Grauer, the co-inventor of Cantometrics, admitted in personal communication (February 2011) that this descriptor is the most difficult to rate.

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

2. PHONATION MODES IN VOICE ACOUSTICS – PREVIOUS WORK

Due to complications in terminology of narrow vs. wide singing within and across disciplines as well as to the subjectivity of the distinction between phonation modes, we turn to voice acoustics for objective definitions and physically measurable effects. The four phonation modes introduced by Sundberg [15]: breathy, neutral, flow (called resonant by other authors) and pressed are vocal production qualities resulting from the voice source (the vibrating vocal folds). In particular they are closely related to glottal resistance which is defined as the quotient of subglottal pressure to glottal airflow. A low subglottal pressure combined with a high glottal flow results in a breathy phonation. Pressed phonation arises when a high subglottal pressure is accompanied by a low glottal flow. The neutral mode lies between these two extremes. The flow phonation is characterised by a lower subglottal pressure than in pressed mode and also by a lower adduction force on the vocal folds. It is an economical voice production mode, because it uses much less effort than in pressed mode gaining a similar sound level, which can be significantly higher than in a neutral mode. At the same time the flow phonation allows various resonances of the vocal tract to be used most effectively, while the pressed phonation tends to restrict some of them.

Given the above, the four phonation modes are not discrete states of vocal production but are rather areas in a continuum space which can be distinguished on the psychoacoustic level. This continuum is not fully ordered: while breathy and pressed modes represent its two extremes, there are endless states between them, and flow phonation is a sweet spot in that continuum which optim-

a higher range often cannot be distinguished from an Arabic singer of a similar range in terms of their phonation mode usage. Also, some Gregorian chant interpreters such as Ensemble Organum deliberately use flow phonation in their performance.

ises psychoacoustic values such as loudness and overtone richness. As we know from singing and teaching practice, a singer is usually capable of using one or more localities of this space.

While the phonation mode of a singing fragment can only be identified subjectively in a psychoacoustic experiment, the glottal wave - the signal produced by the voice source - can be measured during singing by means of a laryngograph (electroglottograph), a non-invasive tool which sends a small current through the larynx and records the changes in resistance [5, 12]. Figure 1 shows typical graphs of the glottal wave in all phonation modes.

Electroglottography can be used to measure the glottal wave during the singing process. If, in contrast, audio recordings of previous events are studied, this technique is not applicable.

To analyse the sound production of the voice source in a recording the technique called inverse filtering is often used: the resonances of the vocal tract are estimated from the original signal and a filter is constructed to eliminate them [2, 4, 8, 19]. Applying this filter to the original signal results in an estimation of the glottal wave.

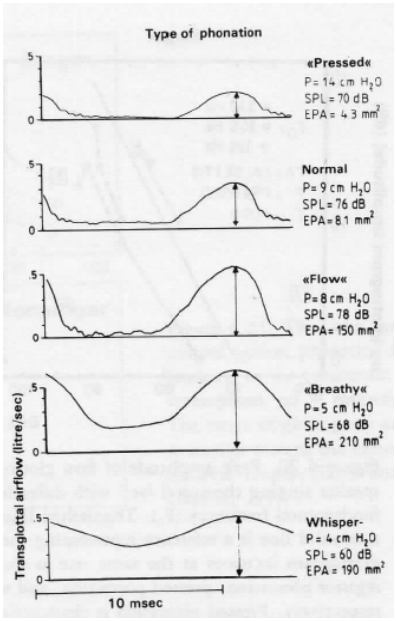

Figure 1. Typical graphs of the glottal wave pulse functions in various phonation modes (from [15], p. 85, used with permission of Northern Illinois University Press)

A number of publications dedicated to detection of pressed and breathy phonation modes employed descriptors derived from the glottal wave such as amplitude quotient (AQ), normalised amplitude quotient (NAQ) and the difference between the first two harmonics (H1-H2) [3, 9, 11, 16, 19].

Unfortunately, all of them rely on internal datasets for their experiments which are neither well documented and controlled nor are they available to other researchers for benchmarking or new studies.

There have been single attempts to determine dominant phonation modes or typical values of glottal wave descriptors for various singing styles. For example Thalén and Sundberg [18] studied Western classical music, pop, jazz and blues, and in a later publication Zangger Borch and Sundberg [1] looked at rock, pop, soul and Swedish dance. Both these studies worked with recordings by just one singer. As a starting point both studies were certainly instructive. Unfortunately it is virtually impossible to make generalisations about a musical style based on samples from just one singer. At the same time the methodology suggested in these papers doesn't scale to batch processing applications. The datasets were not made available to other researchers.

In order to address high-level semantic questions about music like the relationship between a singing style and phonation modes using MIR, new approaches have to be developed that would allow processing of large, real-life data collections. One of the main obstacles for such a development is a lack of reference and training datasets that are well documented and supported and are available to all researchers.

In this paper we present a new dataset that will close this gap and will be a first step in the development of new scalable methods for study of phonation modes in singing. While we also start with recordings by only one singer, the format of the dataset is extensible (see Section 3.4). We plan to add further recordings in future as described in Section 4. Contributions by other researchers will also be welcome.

3. THE DATASET

3.1 The recordings

The dataset consists of ca. 700 WAV files. Each file contains a single recording of a sustained sung vowel. Recordings are of 750 sec length on average. We recommend to use 300 ms around the middle of the samples for analysis - here we can guarantee a relative stability in pitch, intensity, phonation and articulation (beginnings and ends of the samples can be less stable).

Sound	examples	Symbols used in the labels
[a:]	/a/ - low front unrounded sound, like in English ***fa**ther*, German *R**a**t* or in Russian *м**а**м*	A
[e:]	/e/ - high-mid front unrounded vowel, like in English *g**e**t*, German ***E**sel*, Russian *м**е**сто*	E
[i:]	/i/ - high front unrounded, like in	I

	English *fr**ee***, German *Gen**ie***, Russian ***ɐuð***	
[o:]	/o/ - high-mid back rounded, like in German *r**o**t*, Russian ***ком,*** somewhat similar to English *c**au**ght*	O
[ø:]	High-mid front rounded vowel, like German /ö/ in *sch**ö**n*	OE
[u:]	/u/ - high back rounded, like in English *b**oo**t*, German *F**uß***, Russian ***луг***	U
[y:]	High front rounded sound, like German or Turkish /ü/, e.g. in German *m**ü**de*	UE
[ɨ:]	High central unrounded vowel, Russian /ы/ like in ***мы***, similar to English *ros**e**s*	Y
[ɛ:]	Low-mid front unrounded, German /ä/ like in ***Ä**hre*, Russian /э/ like in этот, similar to [æ] in English *c**a**t*	AE

Table 1. The vowels represented in the dataset.

Pitches	Modes
A3 - G4	Breathy, neutral, flow, pressed
G4# - C5	Breathy, neutral, pressed
C5# - G5	Breathy, neutral

Table 2. This table indicates which phonation modes are represented for particular pitches in the dataset.

The vowel sounds represented on the recordings are listed in Table 1. These sounds were sung on all pitches on a semitone scale from A3 to G5, in every phonation mode given in Table 2.

3.2 The singer

All the recordings were produced by one female singer. This excludes any variation that would necessarily arise between singers, which is useful particularly at the initial stages of classification model training and testing.

The singer was professionally trained, with expertise in Western popular and in Russian traditional singing and a profound experience in a number of other music traditions.

The singer's vocal range is approximately D3 – C6, with the working range being usually limited to G3 – F5. At both extreme ends of the range, phonation became unreliable and they were not included into the dataset. The singer's break between the modal and the head (falsetto) register is around E5, thus the surrounding pitches(D5# to F5#) can also be less reliable. Still we decided to include vocalisation in the head register into the dataset to make it more representative, thus all pitches up to G5 were included.

In the head register the singer was unable to produce pressed sounds, thus the pressed phonation mode is only represented up to the upper end of the modal register (see Table 2). Why this is the case seems to be an unsolved problem. While this seems to be common among singers of various traditions in Europe and the Near East, it is unclear whether in other cultures (e.g. in some East Asian traditions) the singers are in fact capable of producing pressed vocalisation in their head register. This observation leads to the question whether the ability to use particular phonation modes on particular pitches is innate or ontogenetic (culturally constructed).

Also, flow phonation could only be produced in the chest voice – up to A4 (recordings up to G4 retained for the dataset). Above A4 it becomes impossible to sing most vowels in the flow mode; at the same time, the neutral mode in the middle and head voice partly gains the qualities of the flow mode, such as intensity and richness in overtones, though it is very different from the chesty flow phonation. The singer reported from her experience of teaching Russian traditional singing, which heavily uses the flow mode, that this limit is typical for female singers, though some exceptional performers are capable of producing the flow phonation at as high as C5.

In the lower range, at G3 and below, the opposite is the case: the neutral phonation becomes more and more similar to the flow mode – for this reason recordings below A3 were excluded from the dataset.

The singer apparently had more difficulties with some vowels than with others in particular modes. For example, high front sounds like [i:] and [y:] proved to be harder to achieve in flow phonation.

3.3 Recording conditions

The recordings were made with Olympus LS10 linear PCM digital recorder. We chose 96 kHz sampling rate and 24 bits bit resolution in compliance with the recommendations for acoustic analysis and archiving by the International Association of Sound- and Audiovisual Archives (IASA TC-04) [6].

The built-in high-sensitivity, low-noise stereo microphone of Olympus LS10 is a combination of two microphone heads positioned at an 90° angle. It has an overall frequency response 20 – 44000 Hz. In the frequency range of 150 -3000 Hz it displays a flat frequency response of ±2dB and in the range up to 20 kHz the response is ±5dB.

The lowest fundamental frequency recorded was 220 Hz (A3) which is about ten times higher than the microphone's low frequency response limit. This guarantees the flat phase response and preserves the exact shape of the

waveform – a necessary condition for applications such as inverse filtering [17].

The highest frequencies perceived by the human ear are about 20 kHz which is within the microphone's flat response range and is way below the half of the upper limit of the microphone's frequency response. See [17] for detailed instructions on the choice and positioning of the microphone.

The recorder and the microphone were positioned horizontally at the level of the singer's mouth, at the distance of 50 cm as recommended by the manufacturer for best voice capturing.

The recording session took place in a quiet room environment. The requirement of a signal-to-noise ratio of at least 15 dB has been adhered to [17].

3.4 The labels

The metadata is stored in a table of a relational database, see Table 3. This way of organising metadata is advantageous, because it can be easily extended by further fields and is scalable for unlimited number of entries and relationships. For example if we add recordings by other singers, a new field indicating the singer will be introduced to the table.

Labels were provided by the singer. They mark the pitch, the vowel and the phonation mode the singer intended to reproduce. We also performed a listening test, excluding all recordings where phonation was ambiguous or of a poor quality.

Metadata fields	ID	File	Pitch	Vowel	Phonation mode	Version
example	212	212.wav	C5#	AE	breathy	2

Table 3. Metadata fields of the dataset. For vowel symbols please consult Table 1. Version is optional, it is only used when several recordings of the same vowel at the same pitch in the same phonation mode were recorded. Currently simple numeric IDs are used. In future a use of content derived IDs is planned.

3.5 Dataset availability and license

The dataset will be made available for download under Creative Commons CC BY-NC-SA license. This license allows free sharing of the dataset as well as altering it or building new work based upon it. There are following conditions for the use of the dataset according to this license:

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

4. FUTURE WORK

Our dataset is a first step in creating reference and training collections for the study of phonation mode use in singing. There are several directions in which this dataset can be improved and extended:

1. To further improve recording quality for the particular task of glottal wave estimation a professional microphone specifically designated for voice measurements should be used for the production of the recordings (LS2-type microphones as specified by IEC 61094-1 and ANSI S1.15-1997 standards). For other experiment designs, a dataset with varying recording quality and recording conditions could be useful.

2. To enhance the reliability of the labels, they can be verified by independent experts, ideally by singing teachers representing various music cultures.

3. Electroglottograph can be used to measure the glottal wave at the singer's glottis during recording. This would allow more objective judgements about the phonation mode. These measurements would also provide an excellent reference for glottal wave estimations on new, unseen data.

4. Alternatively, if an exact measurement of the glottal airflow is required, an airflow mask developed by Rothenberg can be used, which also has an advantage of the low frequency limit of 0 Hz [14].

5. The scope of the dataset can be generalised by including recordings of other singers, male, female as well as children. This would introduce inter-performer variation, which needs to be studied and is necessary to construct real-life classifiers. It is important that singers from different musical traditions are represented, because the ability to utilise various phonation modes can vary greatly across cultures. Ideally, a representative dataset with recordings from all around the globe could be compiled, which would allow to study the distribution of phonation mode use in singing among humans.

6. Another way of generalisation, in particular in view of practical tasks of automatic phonation mode detection, would be to introduce recordings by groups of singers, from small groups to large choirs. Also, recordings where singers are accompanied by musical instruments could be included.

5. CONCLUSIONS

Phonation mode is an important characteristic of singing, playing a vital role in many singing-related disciplines. It remains under-researched, one of the reasons being the lack of reference and training data. The dataset presented here closes this gap. It is aimed at MIR researchers who wish to develop automated methods for phonation mode detection in singing.

6. REFERENCES

[1] Borch, D. Z. and Sundberg, J. (2011). Some phonatory and resonatory characteristics of the rock, pop, soul, and swedish dance band styles of singing. *J Voice*, 25(5):532–7.

[2] Drugman, T., Bozkurt, B., and Dutoit, T. (2012). A comparative study of glottal source estimation techniques. *Computer Speech and Language*, 26:20–34.

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

[4] Fritzell, B. (1992). Inverse filtering. *Journal of Voice*, 6(2):111–114.

[5] Froeschels, E. (1943). Hygiene of the voice. *Arch Otolaryngol.*, 38(2):122–130.

[6] Gudnason, J., Mark R.P. Thomas, D. P. E., and Naylor, P. A. (2012). Data-driven voice source waveform analysis and synthesis. *Speech Communication*, 54:199–211.

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

[8] (2009). *Guidelines on the Production and Preservation of Digital Audio Objects: Standards, Recommended Practices and Strategies (IASA-TC 04)*. IASA (International Association for Sound- and Audiovisual Archives) Technical Committee, 2 edition.

[9] Lehto, L., Airas, M., Björkner, E., Sundberg, J., and Alku, P. (2007). Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *J Voice*, 21(2):138–50.

[10] Lomax, A. (1977). *Cantometrics: A Method of Musical Anthropology (audio-cassettes and handbook)*. Berkeley: University of California Media Extension Center.

[11] Orr, R., Cranen, B., de Jong, F., d'Alessandro, C., and Scherer, K. (2003). An investigation of the parameters derived from the inverse filtering of flow and microphone signals. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL '03)*. Taalwetenschap Otorhinolaryngology.

[12] Pulakka, H. (2005). Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. Master's thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, Department of Computer Science and Engineering.

[13] Ramig, L. O. and Verdolini, K. (1998). Journal of speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, 41:101–116.

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

[15] Sundberg, J. (1987). *The science of the singing voice*. Illinois University Press.

[16] Sundberg, J., Thalén, M., Alku, P., and Vilkman, E. (2004). Estimating perceived phonatory pressedness in singing from flow glottograms. *J Voice*, 18(1):56–62.

[17] Svec, J. G. and Granqvist, S. (2010). Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, 19:356–368.

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

[19] Walker, J. and Murphy, P. (2007). A review of glottal waveform analysis. In *PROGRESS IN NONLINEAR SPEECH PROCESSING*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer.