Does singing style correlate to social behaviour? - A revision of the Cantometric descriptor *vocal tension* and its correlation to the subordination of women in society

POLINA PROUTSKOVA
Goldsmiths, University of London
proutskova@googlemail.com

Abstract

This is a presentation of work-in-progress for the FMA conference 2012 in Seville. The project described in this article is concerned with the Cantometric descriptor *vocal width* or *vocal tension*. This vocal quality was shown to correlate with subordination of women in the society by Alan Lomax and the Cantometrics team. In the current project the choice of this aspect of Cantometric parametrisation is put under scrutiny. Methods for automated extraction of the relevant vocal production qualities are investigated with the goal to reproduce or refute Cantometrics' findings using new data and avoiding the subjectivity of manual rating. Inverse filtering is identified as a possible approach for semi-automatic detection of phonation modes and ways to automate it are discussed.

1. Motivation: *vocal tension* in Cantometrics

The Cantometrics project led by ethnomusicologist Alan Lomax in the 1960s-1990s aimed at finding correlations between singing styles customary in a society and its social traits. Lomax argued that singing being a mode of communication is highly regulated and encapsulates the society's traditions and values; thus it must carry general information about communication modes of the given society (Lomax, 1976).

One of Lomax's most intriguing hypotheses was the correlation between a Cantometric descriptor called *vocal width* and the subordination of women. Lomax suggested that in societies where narrow, squeezed, tense vocalisation is the norm, pre-marital sex is strongly forbidden for women and vice versa, where singing is relaxed and open-throated, the rules regarding the pre-marital behaviour of women are also more relaxed (Lomax, 1968).

He arrived at this conclusion after analysing 5000 music samples from over 400 societies (The Cantometrics Dataset). For each sample 36 Cantometric descriptors were rated by human listeners to create a musical profile of the society. The profiles were then compared with anthropological data on social traits. A statistically significant correlation was found between vocal tension and pre-marital sex norms (Lomax, 1976).

The subordination of women hypothesis has been contested by ethnomusicologists, and Lomax's methodology involving human raters was also criticised. Vocal width is one of the hardest descriptors to rate in the Cantometric system. The consensus among raters was the lowest for this descriptor. The dichotomy between tense vocalisation and open throat is undermined by Russian traditional vocal production, in which a highly (in)tense sound is achieved by singing with a wide, open throat. Also strong dependencies between descriptors like loudness, nasality, rasp, yodel and vocal tension call for a revision of this aspect of Cantometric parametrisation. An empiric method based on acoustic analysis, allowing for an objective and reproducible examination is required.

2. Previous work: phonation modes and their detection

In his classical book "The singing voice" Johan Sundberg identifies four different phonation modes in singing: breathy, neutral, flow (called resonant by other authors) and pressed (Sundberg, 1987). These phonation modes are qualities of vocal production resulting from the voice source (the vibrating vocal folds), in particular from decreased or increased glottal resistance.

To analyse the sound production of the voice source 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 (Fritzell 1992, Walker and Murphy 2007, Drugman et al. 2012, Gudnason et al. 2012). Applying this filter to the original signal results in an estimation of the glottal wave - the signal produced by the glottis.

Many publications dedicated to detection of pressed and breathy phonation modes rely on 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) (Orr et al. 2003, Walker and Murphy 2007, Drugman et al. 2008).

While the phonation mode of a singing fragment can only be identified subjectively in a psychoacoustic experiment, the glottal wave 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 (Howard 2010, Pulakka 2005).

3. Experiment design

The experiment presented here was designed as a pilot to demonstrate that phonation modes can be reliably extracted from conventional audio recordings of sustained sung vowels in a fully automated way. If this can be achieved, a generalisation of the method can be considered to include recordings by several singers, from different cultures, with a varying recording quality.

3.1 Data

A data set of sung sustained vowels in all four phonation modes was recorded by the author, who is an experienced singer with the knowledge of various musical traditions. The dataset includes nine different vowels in each phonation mode covering pitches on a semi-tone scale in the range between C4 and A4. Above A4 only two phonation modes - neutral and breathy – were recorded. The dataset consists of 100 recordings, which were recorded as 128 kB/s MP3 files. Recordings were made using Olympus LS10 digital recorder and its built-in stereo microphone. The distance from microphones to the lips was approximately 35 cm and was kept constant as far as possible by eye control during recording.

This dataset with its phonation mode labels will provide ground truth for the experiment.

This dataset can only be considered as preliminary, because recording conditions are crucially important for accurate inverse filtering. A new dataset of recordings will have to be created, recorded in a non-compressed audio format. A better control of the distance from the source to the microphone will have to be executed by fixing the positions of the microphone and of the singer's head. Also a sound pressure level calibration will have to be performed by playing a generated sine wave of a known frequency in the recording room and measuring its pressure at the recording microphone (Svec and Granqvist, 2010).

3.2 Classification using TKK Aparat

The software package TKK Aparat developed by Matti Airas at Helsinki Institute of Technology (Airas, 2008) implements a semi-automatic algorithm based on Iterative Adaptive Inverse Filtering (IAIF). As opposed to a manual implementation represented by Decap from KTH, Aparat requires manual optimisation of only two parameters: 1. number of formants and 2. lip radiation (Lehto et al., 2007). Given that the software is written in Matlab, it could easily be extended by a module for automatic parameter optimisation.

The goal of this experiment is to investigate two parameter optimisation strategies in TKK Aparat:

1. Optimize parameters for best prediction of the four phonation modes. This can be done as a cross-validation using a classification algorithm such as Support Vector Machines (SVM). In each cross-validation cycle a grid search will be performed for Aparat parameters (number of formants and lip radiation) as well as for SVM parameters C and γ. For each point on the grid and each recording in the training set an estimation of the glottal wave will be produced by

Aparat and AQ, NAQ and H1-H2 descriptors will be calculated. An SVM model will then be trained based on ground truth and these descriptors. The model is then evaluated on the left out test set.

2. Optimize to arrive at the expected contour of the glottal wave graph. In various publications on inverse filtering for speech and singing voice parameter optimisation is described in terms of arriving at the expected form of the glottal wave graph (Lehto et al. 2007, Airas 2008). This form is characterised by a bell-like shape of the pulse with a clear closed phase (except breathy phonation), a rather steep closing phase and a strong negative peak of the derivative. Time-domain descriptors of the glottal wave such as normalised magnitude of the negative peak of the derivative or normalised closed quotient can be used to capture the above qualities of the glottal wave graph. These descriptors can easily be calculated based on the glottal wave descriptors returned by Aparat.

The results of two optimisation strategies will have to be compared and if significant differences arise, they'll have to be analysed and accounted for.

4. Conclusion

If the described experiment is successful, TKK Aparat will be extended by an automation module and a phonation mode detection module. A generalisation of the strategy can be attempted and evaluated on a varied dataset such as the Cantometrics Training Tapes collection. If this can be achieved, an opportunity will present itself to revise the Cantometrics approach and to re-investigate its exciting and speculative findings about the main question of ethnomusicology: the relationship of music and culture.

Acknowledgements

I'd like to thank my supervisors Geraint Wiggins, Christophe Rhodes and Tim Crawford as well as Victor Grauer, the co-inventor of Cantometrics, for their invaluable contribution.

References

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

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)*.

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

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

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.

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

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.

Lomax, A. (1968). *Folk Song Style and Culture*. Transaction Books, New Brunswick, New Jersey.

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

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