# Deep throat as a source of information

*Mattias Heldner[*], Petra Wagner[†‡] and Marcin Włodarczak[*]*
** Department of Linguistics, Stockholm University, Sweden*
*† Faculty of Linguistics and Literary Studies, Bielefeld University, Germany*
*‡ Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden*

***Abstract***
*In this pilot study we explore the signal from an accelerometer placed on the tracheal wall (below the glottis) for obtaining robust voice quality estimates. We investigate cepstral peak prominence smooth, H1-H2 and alpha ratio for distinguishing between breathy, modal and pressed phonation across six (sustained) vowel qualities produced by four speakers and including a systematic variation of pitch. We show that throat signal spectra are unaffected by vocal tract resonances, F0 and speaker variation while retaining sensitivity to voice quality dynamics. We conclude that the throat signal is a promising tool for studying communicative functions of voice prosody in speech communication.*

## Introduction

Voice quality plays an important role in human communication. Voice quality contains information related to the speaker's vocal health (e.g. Maryn & Weenink, 2015; Sauder, Bretl, & Eadie, 2017). It adds to the affective content of what is being said (e.g. Airas & Alku, 2006; Gobl, 2003; Scherer, Sundberg, Tamarit, & Salomão, 2015). There are languages that employ voice quality for making phonemic contrasts (e.g. Gordon & Ladefoged, 2001; Kuang & Keating, 2014). Importantly, voice quality is also relevant for domains larger than single segments and therefore has "prosodic" functions, such as marking word- and utterance-level prominences (e.g. Kakouros, Räsänen, & Alku, 2017; Sluijter & van Heuven, 1996; Yanushevskaya, Chasaide, & Gobl, 2017), boundary phenomena (e.g. Carlson, Hirschberg, & Swerts, 2005) and turn-taking (e.g. Gravano & Hirschberg, 2011; Ogden, 2002). Such observations of suprasegmental functions of voice quality led to coining the term *voice prosody* for studies of the communicative functions of voice (Gobl, Yanushevskaya, & Chasaide, 2015).

Consequently, voice quality dynamics is a relevant topic in speech communication research. A major obstacle in this pursuit, however, is that it is difficult to measure relevant aspects of voice quality in a reliable way, and especially in continuous or conversational speech. There are several reasons for this. First of all, a lot of voice quality research is based on sustained vowels and many established voice quality measures (e.g. jitter and shimmer) require such data in order to be meaningful. Furthermore, voice quality measurements often involve *glottal inverse filtering* techniques to remove the effects of the vocal tract and the lip radiation from the microphone signal (e.g. effects of formants on spectrum slope). While automatic inverse filtering techniques exist (see e.g. Alku, 2011 for a review), they are generally not considered accurate enough when applied to continuous speech (Alku, 2011; Gobl, et al., 2015). Thus, voice quality researchers often resort to manual glottal inverse filtering, which is both very time consuming and requires highly skilled experimenters (Gobl, et al., 2015). As a consequence, voice quality studies on large-scale conversational speech are scarce.

Inspired by recent work using accelerometers (aka throat microphones) placed on the neck surface below the glottis for ambulatory voice monitoring (Mehta, et al., 2015), as well as own recent experiences with throat microphones for capturing breathing noises (Włodarczak & Heldner, 2017), in this pilot study we explore accelerometer signals for obtaining voice quality measures. Thus, the primary goal of this paper is to explore whether accelerometers placed on the tracheal wall are sensitive to voice quality dynamics without the need for inverse filtering of the throat microphone signal (as in Chien, Mehta, Guenason, Zañartu, & Quatieri, 2017; Llico, et al., 2015; Zañartu, Ho, Mehta, Hillman, & Wodicka, 2013). A secondary goal is to evaluate the robustness of the throat microphone signal to formant variation, pitch variation, and pitch level.

While the long-term goal is applying such methods to continuous speech, we take sustained vowels as a starting point here.

# Materials & methods

## Subjects

Three semiprofessional singers (2 females, 1 male) with phonetic expertise and one expert phonetician (1 male) served as voice talents.

## Recording

All recordings took place in a sound treated room at Stockholm University. During recording, participants produced sequences of 6 sustained vowels /aː, eː, iː, yː, uː, oː/ at 4 different pitch levels each, covering one octave, and with 3 different voice qualities (modal, breathy, tense). The recordings were ordered by voice quality, that is participants chose an individually comfortable low pitch level, a vowel and a voice quality to start with, e.g. modal /aː/, and then produced modally voiced sequences for each vowel starting from their low comfort pitch level, then successively raising pitch by a major third until a full octave was reached, and then successively lowering pitch until the base pitch is reached again. The participants produced the same sequence for the remaining voice qualities. Participants were asked to target 1-2 seconds for each sustained vowel, but neither durations nor pitch levels were strictly controlled. Per speaker, each combination of *vowel-pitch-quality* is recorded twice, except for the highest pitch level, and 7 recordings were made for each *vowel-quality* combination. In total, 672 vowels were recorded.

## Data acquisition

The speech signal was recorded using a directional headset microphone (DPA 4088) placed 3 cm from the corner of the mouth. This microphone has a flat frequency response up to 1 kHz and a soft boost (4-6 dB) up to 15 kHz. The throat signal was recorded using a miniature accelerometer (Knowles BU-27135) attached to the skin on the tracheal wall (below the cricoid cartilage) with cosmetic glue (see Figure 1). This accelerometer has a flat frequency response from 20 Hz to 3 kHz and a 4 dB boost up to 6 kHz. We use the same accelerometer as in Mehta, et al. (2015), and the sensor was made in the Phonetics Laboratory at Stockholm University.

*Figure 1. Accelerometer attached to the skin on the tracheal wall.*

Both signals were connected to a Shure ULX-D digital wireless system and recorded using the REAPER software.

## Acoustic measures

We captured three aspects of voice quality: (i) signal periodicity, (ii) the relative amplitude of the first harmonic, and (iii) spectral tilt.

The signal periodicity was assessed by Cepstral Peak Prominence (CPP, Hillenbrand, Cleveland, & Erickson, 1994). Defined as the amplitude of the first peak in the real cepstrum (first rahmonic) of a sound, relative to the cepstrum trend line, CPP has been used extensively in clinical literature as a measure of dysphonia (Sauder, et al., 2017). It also been successfully used for detection of breathiness and, with somewhat mixed results, for assessment of the overall voice quality (see Fraile & Godino-Llorente, 2014 for a review). In this paper, we used the smoothed version of CPP (CPPS, Hillenbrand & Houde, 1996), following the procedure outlined in Watts, Awan, and Maryn (2017).

The relative amplitude of the first harmonic (i.e. the fundamental) was measured using H1-H2, which is a measure of the amplitude of the first harmonic in dB relative to the second harmonic (Hillenbrand & Houde, 1996). Note however, that H1-H2 can also be viewed as a measure of spectral tilt (in dB per octave) in the lower part of the spectrum (cf. Kakouros, et al., 2017; Titze & Sundberg, 1992).

Spectral tilt was measured using the alpha ratio (Frokjaer-Jensen & Prytz, 1976), which is a measure of spectral balance, defined as a ratio of energy below and above 1000 Hz.

Each of the measures was calculated for the speech signal as well as for the throat signal. H1-

H2 was additionally calculated for an estimation of the voice source in the speech signal obtained using an automatic inverse filtering method (Airas, 2008; Alku, 1992). All measures were *z*–normalized by speaker and microphone.

We have used freely available speech processing tools for all of the analyses in this paper. CPPS and alpha ratio were calculated in Praat (Boersma & Weenink, 2018), H1-H2 was obtained from the COVAREP repository (Degottex, Kane, Drugman, Raitio, & Scherer, 2014). All features were subsequently speaker-normalized. Additionally, since H1-H2 is likely to be affected by speaker's F0, we split values of these features of the median F0 calculated for all speakers (186 Hz).

### Analyses

For this pilot study, we restricted the analyses to (i) qualitative descriptions or illustrations of why the throat signal may provide a more robust estimation of voice quality and (ii) descriptive statistics of the acoustic measures to allow a comparison of how well the different measures separate the different voice qualities.

## Results

A first illustration of why the throat signal may potentially be useful for voice quality measures is given in Figure 2 showing LPC spectra of the same vowel from a microphone signal, a throat signal and an inverse filtered microphone signal using an LPC based inverse filter function in Praat. It is easy to see that vocal tract formants will influence any microphone-based characterization of spectral tilt involving the F1 to F3 frequency region. In contrast, there is no evident influence of vowel formants in the throat signal, although resonances that most likely originate from the subglottal system are visible at approximately 550, 1400 and 2700 Hz (cf. Sundberg, Scherer, Hess, Muller, & Granqvist, 2013).

It is also evident from Figure 2 that the throat signal spectrum is different from the voice source spectrum estimated using inverse filtering. In particular, the throat signal spectrum has an elbow at the first subglottal resonance, whereas the inverse-filtered signal rolls off monotonously. Thus, the throat spectrum is perhaps better characterized by a two-segment slope below and above the first subglottal resonance, or by a polynomial function.

A second, and perhaps more convincing illustration of the benefits of the throat signal, is provided in Figure 3 showing LPC spectra of the throat signals for three different vowels by the same speaker. The three spectra are virtually identical. This indicates that the throat signal is robust to variations in formant frequencies. Similar analyses with different speakers (Figure 4) and with different F0 levels (Figure 5) show that the throat signal is also robust to speaker and pitch variation.

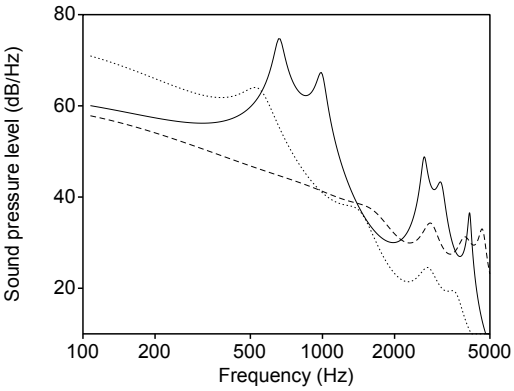


*Figure 2. LPC spectra of the vowel* /a:/ *from the normal microphone signal (solid line), the throat signal (dotted line), and an inverse filtered microphone signal (dashed line). The vowel was produced in modal voice quality by a male speaker (F0 ≈ 115 Hz).*

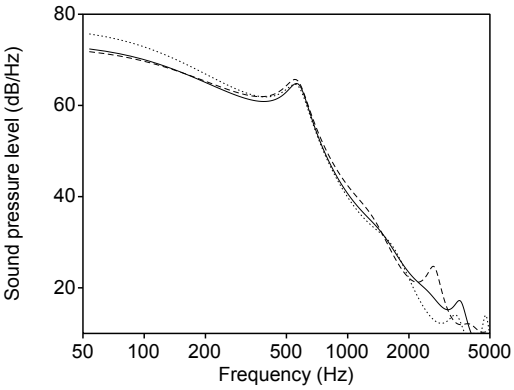


*Figure 3. LPC spectra of three different vowels (*/a:/ *solid line,* /i:/ *dotted line,* /u:/ *dashed line) produced in modal voice quality by a male speaker (F0 ≈ 150 Hz).*

But of course, it is not enough for a signal to be robust to various influences in order to be useful for voice quality measures. It has to be sensitive to relevant voice quality variation as well. Figure 6 illustrates this aspect of throat signals with LPC spectra of different voice qualities (same vowel, same speaker).

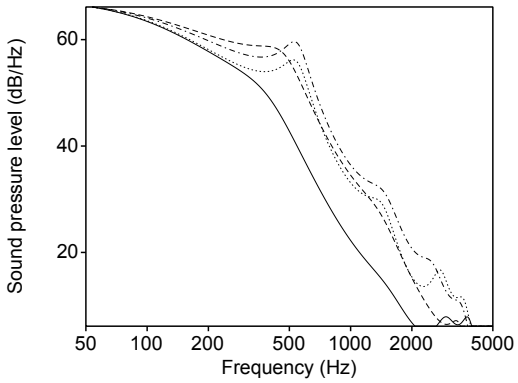


*Figure 4. LPC spectra of modal voice /a:/ by four speakers (2f, 2m). For comparison, the spectra have been shifted on the y-axis.*

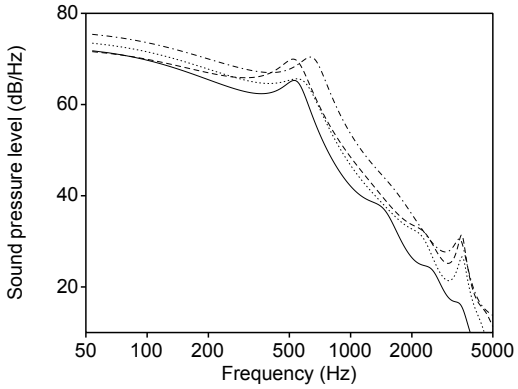


*Figure 5. LPC spectra of the vowel /a:/ at four different F0 levels. Female speaker.*

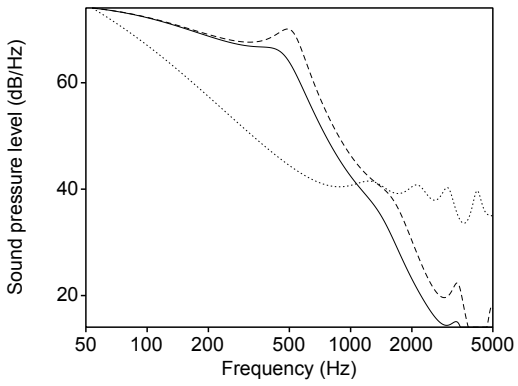


*Figure 6. LPC spectra of three voice qualities (modal: solid line, tense: dotted line, breathy: dashed line) produced on the same vowel* /a:/ *by a female speaker (F0 ≈ 150 Hz). For comparison, the spectra have been shifted on the y-axis.*

These conclusions are further confirmed by the results in Figures 7–9, which show that the throat signal provides a robust separation between the three voice qualities regardless of the measure used. In Figure 7, we plot CPPS values calculated from the throat and speech signals. Overall, the signal periodicity increases from breathy to modal to pressed. Not surprisingly given its original purpose, CPPS is very effective at distinguishing breathy and non-breathy phonations in both types of signals. Additionally, in the throat signal it also offers better separation of modal and pressed phonations.

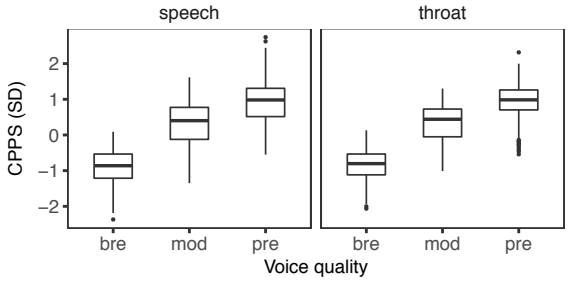


*Figure 7. Box plots of normalized signal periodicity (CPPS) in breathy (bre), modal (mod) and pressed (pre) voice quality.*

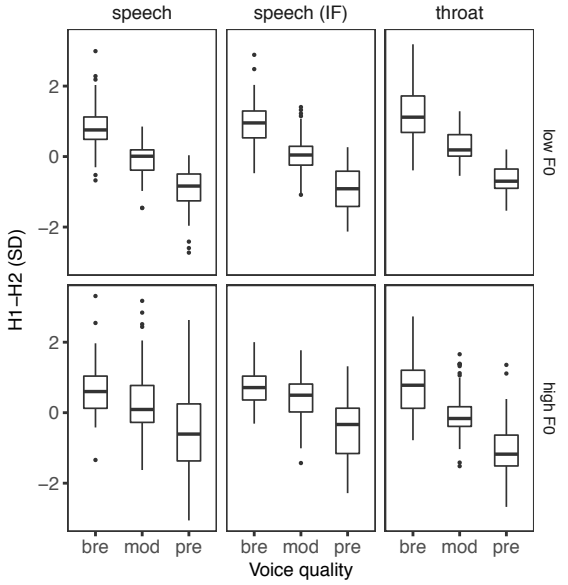


*Figure 8. Box plots of normalized relative amplitude of the first harmonic (H1-H2) in breathy (bre), modal (mod) and pressed (pre) voice quality. Data is split by F0 level with the values below median in the top row and above median in the bottom row.*

H1-H2 (Figure 8), reflecting both relative amplitude of the first harmonic and spectral tilt in the lower part of the spectrum, shows large dependence on F0 level when calculated on the speech signal. Namely, it separates the three voice qualities rather well at low F0 levels but fails for higher F0 values (especially for the breathy-modal contrast). This is most likely due to the fact that for higher pitches the first two har-

monics are increasingly influenced by F1. Notably, this effect is also observed in the automatically inverse-filtered signal, suggesting that residuals of vocal tract resonances must be present in the signal. By contrast, the throat signal is virtually unaffected by fundamental frequency.

Finally, spectral balance captured by alpha ratio (Figure 9) has little discriminatory value when calculated on the speech signal but preserves good separation in the throat signal.

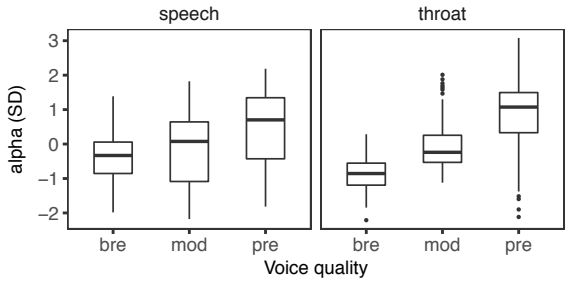


*Figure 9. Box plots of normalized spectral tilt (alpha ratio) in in breathy (bre), modal (mod) and pressed (pre) voice quality.*

# Discussion

Based on the results in the previous section, we conclude that the throat signal is robust to vowel quality, speaker and pitch variation. At the same time, it is sensitive to differences in voice quality. This is the case even though the throat signal is spectrally different from signals obtained using inverse filtering of speech signals (prevalent in voice quality research). Given its robustness and stability, we speculate that we can also eliminate the need for inverse filtering of the throat signal (e.g. Zañartu, et al., 2013) for the purpose of studying the communicative function of voice quality dynamics.

In future work, we will investigate whether the the throat signal is equally useful for studies of continuous, spontaneous or conversational speech. We will also monitor sound pressure level (SPL) given the known dependency between SPL and spectral tilt (e.g. Sundberg & Nordenberg, 2006). Finally, we are planning to evaluate other voice quality measures. In particular, we hope to obtain a better estimate of spectral tilt by using a DNN-based approach (Jokinen & Alku, 2017; Kakouros, et al., 2017), and to explore measures of pitch-strength (Eddins, Anand, Camacho, & Shrivastav, 2016) as an alternative to H1-H2.

In conclusion, the throat signal is a promising, tool for studying communicative functions of voice prosody in speech communication. It could potentially allow quantitative analyses of large speech materials without relying on the error prone automatic inverse filtering methods of speech signals.

# Acknowledgements

This work was partly funded by a Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to the first author; by a Humbolt stipend within the Swedish-German Programme *Research Awards for Scientific Cooperation* to the second author; and by a Christian Benoît Award to the third author.

# References

Airas M (2008) TKK Aparat: an environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology* 33: 49-64. doi: 10.1080/14015430701855333.

Airas M and Alku P (2006) Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica* 63: 26-46. doi: 10.1159/000091405.

Alku P (1992) Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication* 11: 109-118. doi: 10.1016/0167-6393(92)90005-r.

Alku P (2011) Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana* 36: 623-650. doi: 10.1007/s12046-011-0041-5.

Boersma P and Weenink D. (2018). Praat: doing phonetics by computer [Computer program] (Version 6.0.39). Retrieved from http://www.praat.org/

Carlson R, Hirschberg J and Swerts M (2005) Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication* 46: 326-333. doi: 10.1016/j.specom.2005.02.013.

Chien Y-R, Mehta D D, Guenason J, Zañartu M and Quatieri T F (2017) Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25: 1718-1730. doi: 10.1109/taslp.2017.2714839.

Degottex G, Kane J, Drugman T, Raitio T and Scherer S. (2014). COVAREP - A collaborative voice analysis repository for speech technologies *Proc. ICASSP 2014* (pp. 960-964). Florence, Italy.

Eddins D A, Anand S, Camacho A and Shrivastav R (2016) Modeling of breathy voice quality using pitch-strength estimates. *Journal of Voice* 30: 774 e771-774 e777. doi: 10.1016/j.jvoice.2015.11.016.

Fraile R and Godino-Llorente J I (2014) Cepstral peak prominence: A comprehensive analysis. *Biomedical*

*Signal Processing and Control* 14: 42-54. doi: 10.1016/j.bspc.2014.07.001.

Frokjaer-Jensen B and Prytz S (1976) Registration of voice quality. *Brüel and Kjær Technical Review* 3: 3-17.

Gobl C (2003) The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40: 189-212. doi: 10.1016/s0167-6393(02)00082-1.

Gobl C, Yanushevskaya I and Chasaide A N (2015) The relationship between voice source parameters and the maxima dispersion quotient (MDQ). In *Proc. Interspeech 2015*. Dresden, Germany, 2337-2341.

Gordon M and Ladefoged P (2001) Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29: 383-406. doi: 10.1006/jpho.2001.0147.

Gravano A and Hirschberg J (2011) Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25: 601-634. doi: 10.1016/j.csl.2010.10.003.

Hillenbrand J, Cleveland R A and Erickson R L (1994) Acoustic Correlates of Breathy Vocal Quality. *Journal of Speech Language and Hearing Research* 37. doi: 10.1044/jshr.3704.769.

Hillenbrand J and Houde R A (1996) Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. *Journal of Speech Language and Hearing Research* 39. doi: 10.1044/jshr.3902.311.

Jokinen E and Alku P (2017) Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network. *Journal of the Acoustical Society of America* 141: EL327. doi: 10.1121/1.4979162.

Kakouros S, Räsänen O and Alku P (2017) Evaluation of spectral tilt measures for sentence prominence under different noise conditions. In *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 3211-3215. doi: 10.21437/Interspeech.2017-1237.

Kuang J and Keating P (2014) Vocal fold vibratory patterns in tense versus lax phonation contrasts. *Journal of the Acoustical Society of America* 136: 2784-2797. doi: 10.1121/1.4896462.

Llico A F, Zañartu M, Gonzalez A J, Wodicka G R, Mehta D D, Van Stan J H, et al. (2015) Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *Journal of the Acoustical Society of America* 138: EL14-19. doi: 10.1121/1.4922364.

Maryn Y and Weenink D (2015) Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice* 29: 35-43. doi: 10.1016/j.jvoice.2014.06.015.

Mehta D D, Van Stan J H, Zañartu M, Ghassemi M, Guttag J V, Espinoza V M, et al. (2015) Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology* 3: 155. doi: 10.3389/fbioe.2015.00155.

Ogden R (2002) Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31. doi: 10.1017/s0025100301001116.

Sauder C, Bretl M and Eadie T (2017) Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV). *Journal of Voice* 31: 557-566. doi: 10.1016/j.jvoice.2017.01.006.

Scherer K R, Sundberg J, Tamarit L and Salomão G L (2015) Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language* 29: 218-235. doi: 10.1016/j.csl.2013.10.002.

Sluijter A M C and van Heuven V J (1996) Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100: 2471-2485. doi: 10.1121/1.417955.

Sundberg J and Nordenberg M (2006) Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America* 120: 453-457. doi: 10.1121/1.2208451.

Sundberg J, Scherer R, Hess M, Muller F and Granqvist S (2013) Subglottal pressure oscillations accompanying phonation. *Journal of Voice* 27: 411-421. doi: 10.1016/j.jvoice.2013.03.006.

Titze I R and Sundberg J (1992) Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America* 91: 2936-2946. doi: 10.1121/1.402929.

Watts C R, Awan S N and Maryn Y (2017) A comparison of cepstral peak prominence measures from two acoustic analysis programs. *Journal of Voice* 31: 387 e381-387 e310. doi: 10.1016/j.jvoice.2016.09.012.

Włodarczak M and Heldner M (2017) Capturing respiratory sounds with throat microphones. In: J Eggesbø Abrahamsen, J Koreman & W A van Dommelen, eds, *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016*. Frankfurt am Main, Germany: Peter Lang, 181-190.

Yanushevskaya I, Chasaide A N and Gobl C (2017) Cross-speaker variation in voice source correlates of focus and deaccentuation. In *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 1034-1038. doi: 10.21437/Interspeech.2017-1535.

Zañartu M, Ho J C, Mehta D D, Hillman R E and Wodicka G R (2013) Subglottal Impedance-Based Inverse Filtering of Voiced Sounds Using Neck Surface Acceleration. *IEEE Transacions on Audio, Speech, and Language Processing* 21: 1929-1939. doi: 10.1109/TASL.2013.2263138.