A SPECTRAL METHOD FOR ESTIMATION OF THE VOICE SPEED QUOTIENT AND EVALUATION USING ELECTROGLOTTOGRAPHY

Nicolas Sturmel [*]*, Christophe d'Alessandro, Boris Doval*

LIMSI-CNRS
BP 133 F-91403 Orsay, France
{sturmel, cda, boris.doval}@limsi.fr

ABSTRACT

A new method for estimation of the voice speed quotient (S_q) from acoustic signals is presented. The method is based on source filter decomposition using a new signal representation, the Zeros of Z Transform representation.

A source dominated spectrum is obtained using the ZZT decomposition, and then the glottal formant frequency is estimated. The spectral theory of the voice source shows that the glottal formant frequency depends on the fundamental frequency (F_0), the voice open quotient (O_q) and S_q. Using an electroglottographic (EGG) reference for estimation of F_0 and O_q ; S_q can be obtained from the glottal formant frequency.

The estimation algorithm has been implemented and then tested on a database of male and female speech containing EGG and acoustic signals. Three speakers produced 71 vowels under various conditions of vocal effort, tenseness and fundamental frequency. In most of the case, speed quotient estimation gives accurate values, mainly situated between 1.5 and 4. However, in some situations (high F_0, low first vocalic formant, high O_q) the measurements fail and some post-processing would be necessary. Moreover, it seems that this decomposition method could also be used for O_q estimation, considering typical values of S_q and the perceptual just noticeable differences on O_q (about 15 %).

[*]ENS Cachan, France

Index Terms— speed quotient estimation, asymmetry coefficient, EGG, ZZT, source-filter decomposition, voice quality

1. INTRODUCTION

Glottal flow models like the LF model [1] are useful tools for speech and voice signal analysis and signal processing. Most models are time-domain representations of the glottal flow described by a set of five parameters:

1. A_v: peak amplitude of the glottal flow, or amplitude of voicing. The maximum excitation E, defined as the negative peak amplitude of the glottal flow derivative, can be used alternatively;

2. T_0: fundamental period (inverse of F0);

3. O_q: open quotient, defined as the ratio between the glottal open time and the fundamental period. This quotient also defines the glottal closure instant at time O_qT_0;

4. α_m: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient also defines the instant T_m of maximum glottal flow, relative to T_0 and O_q ($T_m = \alpha_m O_q T_0$). The speed quotient S_q, defined as the ratio between open-

ing and closing times, is equivalent to α_m as: $S_q = \frac{\alpha_m}{1-\alpha_m}$;

5. Q_a: the return phase quotient, defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure), and the closed phase duration. In case of abrupt closure $Q_a = 0$.

Several methods have been proposed since a long time for estimation of F_0, A_v and E. Estimation of the other parameters is by no mean straightforward when only the acoustic signal is available. On the one hand parameters estimation often requires source-filter deconvolution by inverse filtering, a challenging problem for signal processing. On the other hand, time domain parameter estimation is difficult and generally lacks robustness. For the estimation of open quotient, one can take advantage of simultaneous acoustic and ElectroGlotto-Graphic (EGG) recordings. Glottal opening and closure instants can be estimated with reasonable precision and robustness on the EGG signal. Robust and automatic methods for estimation of the two last parameters, S_q and Q_a are still challenging voice signal processing. Estimation of O_q and S_q are important because these parameters correlate well with the lax/tense dimension in voice perception. S_q represents mainly the speed of vocal fold closure, which seems an indication of tenseness. To the best of the authors' knowledge, no estimation method for S_q seems currently available. Then it seemed important to work on such a method, and the present paper is focusing on S_q estimation.

Previous studies such as [2] demonstrated that O_q and S_q are often showing a high degree of covariation. In the spectral domain, one can show that O_q and S_q are influencing the main frequency maximum due to the voice source, the "glottal formant". More precisely, the glottal formant frequency, hereafter noted F_g, is a function of O_q, F_0 and S_q. The main idea of this paper is then to reach the value of S_q by estimating O_q, F_0 and F_g, and then by using the known relationship between S_q and those parameters. For the estimation of F_0 and O_q, we will take advantage of EGG recordings [3]. The F_g value will be estimated on the source component, obtained using a new method for source-filter decomposition, the Zeros of the Z-Transform representation (ZZT) [4]. This spectral method is well fitted to estimation of the glottal formant.

The paper is organized as follows. Next section deals with the glottal flow spectrum, and the estimation of F_g, O_q and $F0$. Section 3 describes the algorithm implemented for S_q estimation. Section 4 presents the experimental results obtained. Section 5 discusses the results obtained and proposes some conclusions.

2. ESTIMATION OF O_Q, $F0$ AND F_G

Following the spectral approach presented in [2], the glottal source spectrum is characterized by a maximum on the amplitude spectrum. It can be shown that both the position of this spectral peak (the glottal formant) and its bandwidth are depending mostly on three main parameters: O_q, S_q and F_0. Figure 1 illustrates the dependency of the source spectrum on the parameter S_q. It displays the source spectrum and the glottal formant position for various values of S_q, where O_q, F_0 and E are fixed, and for the LF model.

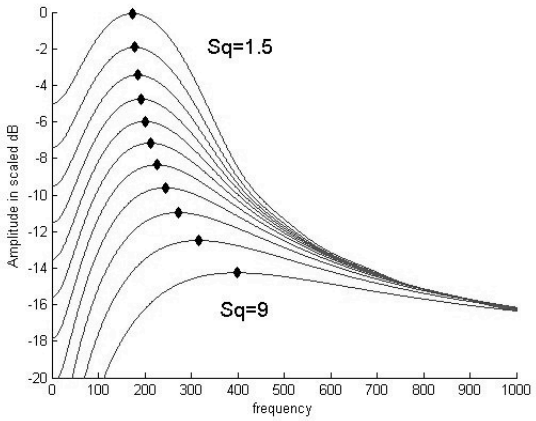

Fig. 1. Glottal source spectrum (curves) and glottal formant variation (diamonds) with S_q varying form 1.5 to 9.

Moreover, it can be shown [2] that for glottal

flow models, the position of the glottal formant can be determined by an equation like: $F_g = f(O_q, F_0, S_q)$. Then the values of S_q can be obtained by estimating the values of F_0, O_q and F_g and using this equation. Let's see now how to estimate these parameters.

2.1. Source/tract decomposition

In order to estimate F_g, one needs some kind of source/filter deconvolution. Here, we have chosen to use the ZZT decomposition method as it is well fitted to glottal formant estimation [5].

The method is illustrated on figure 2. By computing the roots of a Z polynom (top right) whose coefficients are the samples of a two period speech signal (top left), we can perform a causal/anticausal decomposition from the position of those roots in the complex domain. As shown in [6], the source signal (middle left), viewed from the Glottal Closing Instant (GCI) is an anticausal part of the speech signal, whereas the vocal tract response (bottom left) can be viewed as a causal response.

An estimation of F_g is obtained from the source spectrum (middle right) by determination of the amplitude maximum. However, it should be pointed out that the ZZT analysis is very sensitive to the position of the analysis window which should be centered quite precisely at the GCI.

2.2. EGG measurements

An accurate and and reliable value for O_q is mandatory for S_q estimation. Estimation techniques based on the EGG have been proved to be reliable and robust. Then we used EGG recordings and processed them using the DECOM analysis described in [3] to get accurate estimations of O_q and $F0$. EGG is also useful for obtaining reliable GCIs, that are critically needed for the ZZT analysis. Then a simple alignment procedure between the EGG and acoustic signals allows for correct positioning of the analysis windows for the ZZT.

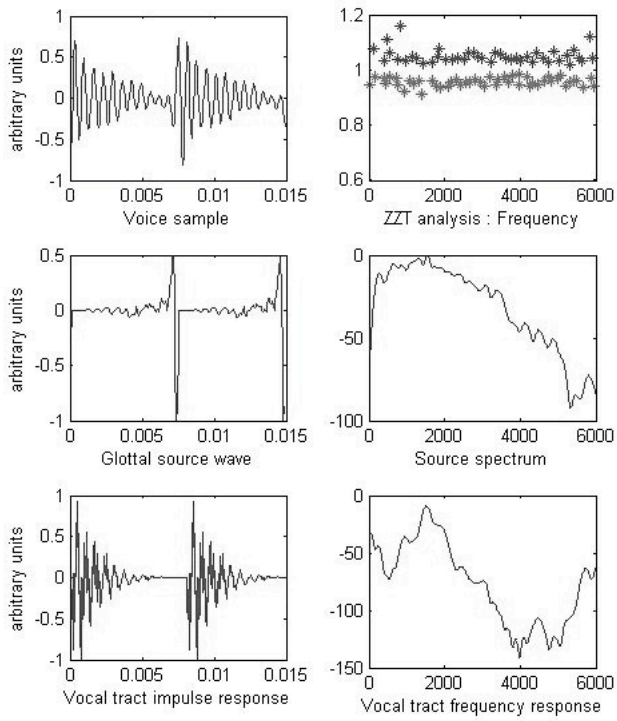

Fig. 2. Example of source-filter decomposition using ZZT. Top left panel: speech signal (vowel /a/, 2 periods). Top right panel: corresponding set of zeros in the spectrum. Middle panels: estimated source waveform (left) and spectrum (right). Bottom panels: estimated vocal tract impulse response (left) and spectrum (right).

3. ALGORITHM IMPLEMENTATION

The algorithm for S_q estimation is displayed in Figure 3. This algorithm contains the 7 following steps:

1. Simultaneous recordings of both the acoustic and EGG signals.

2. The EGG signal is processed for estimation of the GCI, O_q and F_0.

3. The acoustic signal is decomposed into source and filter components. GCI are used for Pitch synchronous ZZT decomposition (equivalent to an inverse filtering in the Z domain). A

source dominated spectrum is obtained together with a vocal tract spectrum (not used here).

4. Then, a local maximum of the source dominated spectrum is searched for in the range $[0.8F_0; 4F_0]$. This gives an estimate of the glottal formant. The next steps of the algorithm are displayed in Figure 4.

5. Glottal flow waveforms (using the LF model) with the estimated O_q and F_0 are synthesized for values of S_q between 1.4 and 4, like in Figure 1.

6. ZZT and glottal formant analysis are performed on these synthetic glottal flow signals. Glottal formants are estimated.

7. The glottal formant estimated on the speech signal and on synthetic LF model waveforms are compared. The closest value of glottal formant gives the estimate for S_q.

Robust and accurate estimation of the glottal formant using the ZZT decomposition method seemed difficult because of a bias mainly due to an estimation error of the D.C. signal component. As this bias seemed systematic but difficult to measure directly, a variation procedure was used for S_q estimation : all possible S_q values were computed and their corresponding F_g frequency measured using ZZT. The corresponding box is detailed on figure 4. The synthetic LF waveforms are computed using the estimated O_q and F_0 parameters. The typical range of S_q values was from 1 to 20, using a logarithmic scale of 30 steps (based on the just noticeable difference measured in [7]). As the LF model is defined by 5 parameters, one has to set the two remaining parameters (amplitude and spectral tilt). However they have little influence on the estimation procedure: a global amplitude variation will not affect at all the peak frequency and a spectral tilt variation hardly changes the glottal formant frequency especially on pressed voiced. Moreover it must be noticed that the source dominated part obtained by the ZZT decomposition corresponds to

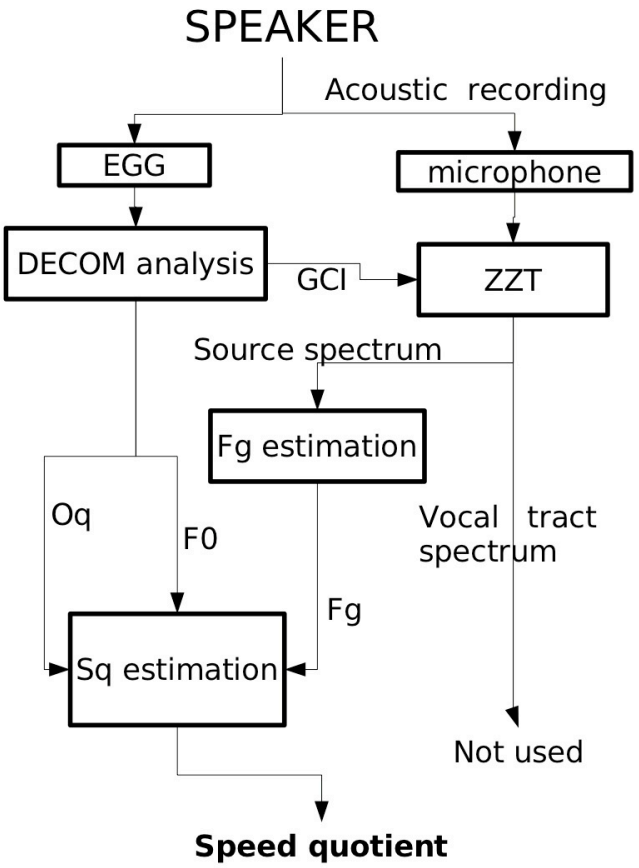

Fig. 3. Algorithm implementation. (*see text for explications*)

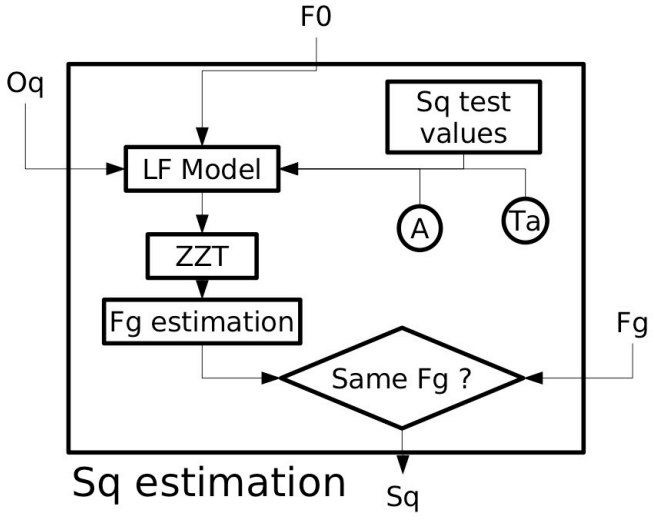

Fig. 4. Zoom on S_q estimation. (*see text for explications*)

the open phase alone so that the return phase (responsible for the spectral tilt) is not present in the estimated source. Then A_v is arbitrarily set to one and T_a to zero, so that the synthesized spectrum is only the spectrum of a glottal open phase (no spectral tilt). Those synthetic glottal waves are then analyzed through ZZT in order to reproduce the same bias in F_g position. At this stage a series of candidate spectrum are produced and compared to the reference signal spectrum. The value of S_q retained by the algorithm is the value corresponding to the closest F_g.

4. EXPERIMENTS

The algorithm was implemented in Matlab. A database of speech and EGG signals was recorded for testing purposes. This database contained 71 speech utterances produced by 2 males and a female speaker. The three cardinal vowels /a/, /i/ and /u/ were uttered with much variation of vocal effort and stress. On the contrary fundamental frequency was kept as constant as possible.

Implementation of ZZT decomposition must be carefully designed. The results were often strongly affected by higher frequency zeros in the spectrum. In some case these perturbations rendered the spectrum difficult to interpret or not readable at all. However, when higher frequency zeros are properly estimated, the source-filter decomposition appeared quite successful. An example is displayed in Figure 5.

On the whole database, about 50% of the utterances were successfully processed by the algorithm. The "good" and "bad" situations are discussed below. Figure 6 shows five utterances of the vowel /a/ (female speaker) with alternatively stressed or relaxed voice quality (average F0=243Hz). It seems that O_q (middle panel) gives a good picture of the underlying voice pressure. S_q is displayed in the bottom panel. S_q is generally low and almost constant for all the utterances. A possible explanation is that the speaker didn't change much vocal effort among the pressed/relaxed utterances. Then only O_q changed and not S_q. Indeed, the SPL is almost constant for all these utterances, another indication of comparable S_q.

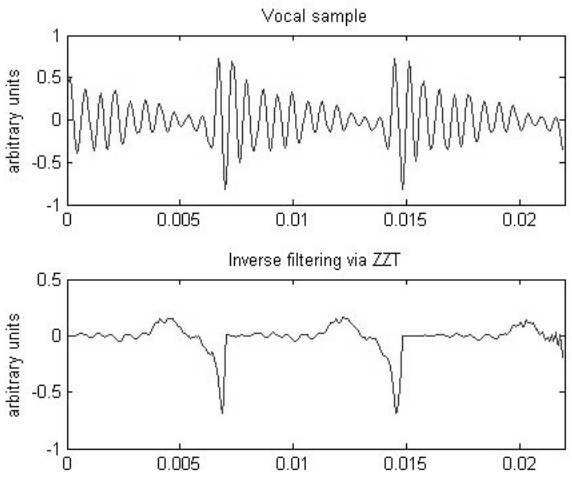

Fig. 5. An application of the ZZT for inverse filtering purposes. Top: original voice sample. Bottom: glottal waveform obtained by ZZT.

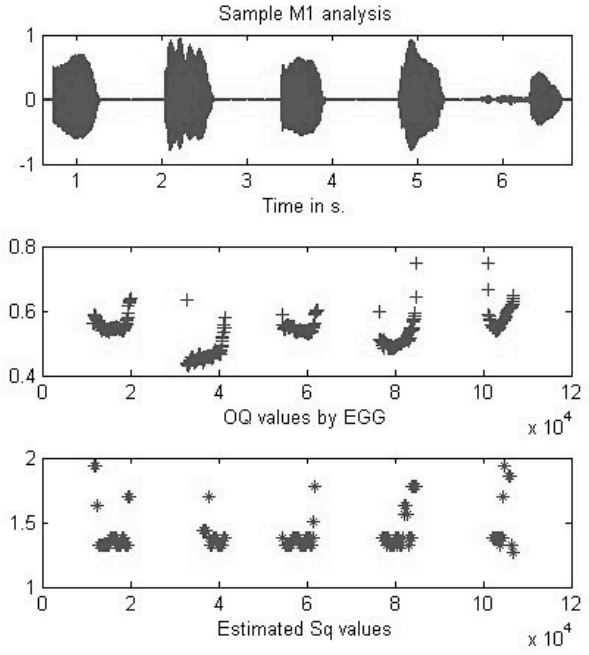

Fig. 6. Speech sample analysis of a female speaker (vowel /a/, 243Hz). Top: original speech sample. Middle: measured O_q values, via EGG. Bottom: estimated S_q values.

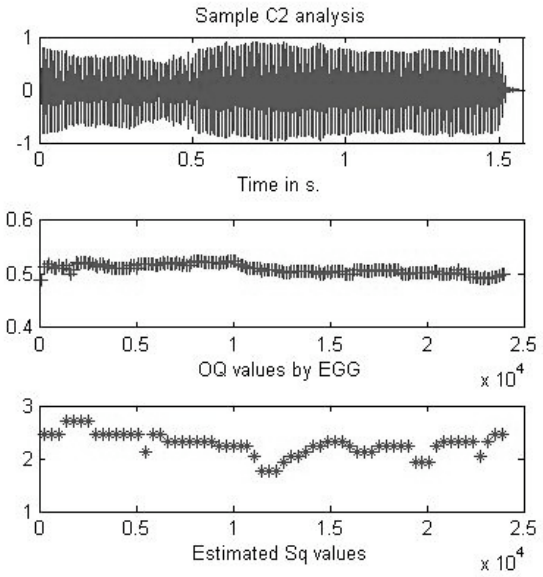

Fig. 7. Speech sample analysis of a male speaker (vowel /a/, 141Hz, stressed voice). Top: original speech sample. Middle: measured O_q values, via EGG. Bottom: estimated S_q values.

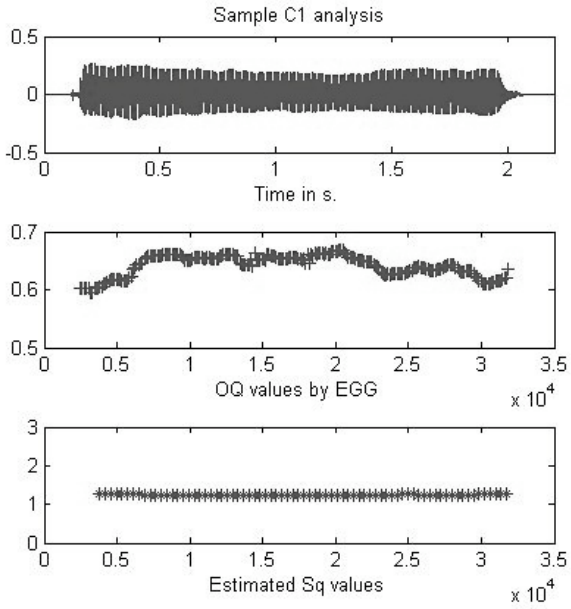

Fig. 8. Speech sample analysis of a male speaker (vowel /a/, 128Hz, relaxed voice). Top: original speech sample. Middle: measured O_q values, via EGG. Bottom: estimated S_q values.

Figure 7 displays an example with more variation of S_q. This utterance is a vowel /a/ produced by a male speaker at an average frequency of 141Hz. The O_q values are low, indicating a pressed voice quality. S_q is quite high (about 3.5) showing that the glottal waveform is rather dissymmetric. It seems that vocal effort and voice pressure are high in this utterance. In contrast to Figure 7, Figure 8 is an utterance with a relaxed voice quality (same male speaker, vowel /a/, average F0=128Hz). This relaxed voice quality corresponds to a higher O_q. S_q is also lower, and indication of a more symmetrical glottal waveform.

5. DISCUSSION AND CONCLUSION

It must be pointed out that Oq and Sq are in principle independent parameters, but that they are often correlated. However, the open quotient represents mostly the pressed/relaxed voice quality, which is independent of vocal effort (a pressed voice can be produced with high or low vocal effort, if vocal effort stands more for spectral tilt and high flow). Sq is also partly correlated with vocal effort (a waveform must be dissymmetric when the vocal effort is high) and partly with Oq. Our preliminary results seem to indicate that different subjects have different settings for Oq and Sq. It seems also that open quotient is more important a parameter than Sq, in practice. For typical Sq variation, the corresponding of Fg is less than 15%. Then, the corresponding variation on open quotient determined by a spectral decomposition is also less than 15%. This variation happens to be lower than the just noticeable difference on Oq according to [7]. Therefore, even if Sq is not taken into account, spectral estimation of Oq using ZZT gives an error within the perceptual bounds for Oq estimation. It could also be convenient to use the amplitude maximum of the source spectrum for S_q estimation. Figure 1 shows that not only the formant frequency is varying along with S_q but also its amplitude A_g. As there could be some problem in scaling appropriately the source spec-

trum from any decomposition method, the variation of A_g with S_q, whose magnitude seems far more important, should then possibly lead to a more accurate estimation method.

Overall Sq estimation proved to be efficient on about one half of the utterances in our database. The analysis conditions leading to successful analyses were: low fundamental frequency, modal register (laryngeal mechanism I), and low values of Oq. Typical estimated values for Sq were mainly between 1.4 (the theoretical minimum for the LF glottal source wave model) and 4.

In summary, we showed how the use of both EGG recordings and source/tract decomposition data could be combined to perform an estimation of the speed quotient, one of the 5 parameters of the common glottal source models. We chose to use the ZZT for source/tract decomposition as it is a simplest way and efficient analysis method for the glottal formant estimation. After a description of the algorithm, we presented some results of Oq and Sq estimation. The results, only relevant for "good" situations such as low fundamental frequency, and low Oq values, showed estimated Sq values consistent with theoretical expectations. Sq seems to be correlated both with the pressed voice quality and vocal effort. There is also an indication that spectral analysis could be used for Oq estimation on an acoustic signal (without EGG) within perceptual bounds.

6. REFERENCES

[1] G. Fant, "The lf-model revisited," *STL-QPSR*, vol. 2-3, pp. 119–156, 1995.

[2] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *to be published in acta acustica*, 2006.

[3] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of derivative electroglottographic signals for characterization of nonpatological phonation," *JASA*, vol. 115 (3), pp. 1321–1332, 03 2004.

[4] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform reprsentation with application to source-filter sparation in speech," *IEEE*, vol. 12, pp. 344–347, 2004.

[5] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "A method for glottal formant frequency estimation," *icslp*, 2004.

[6] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," *VOQUAL'03, Geneva*, 08 2003.

[7] N. Henrich, Gunila Sundin, Daniel Ambroise, C. d'Alessandro, B. Doval, and M. Castellengo, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *JOV03*, vol. 17, pp. 481–494, 2003.