FUNCTION OF PHASE-DISTORTION FOR GLOTTAL MODEL ESTIMATION

Gilles Degottex, Axel Roebel, Xavier Rodet

IRCAM - CNRS-UMR9912-STMS, Analysis-Synthesis Team
1, place Igor Stravinsky, 75004 Paris

ABSTRACT

In voice analysis, the parameters estimation of a glottal model, an analytic description of the deterministic component of the glottal source, is a challenging question to assess voice quality in clinical use or to model voice production for speech transformation and synthesis using *a priori* constraints. In this paper, we first describe the Function of Phase-Distortion (FPD) which allows to characterize the shape of the periodic pulses of the glottal source independently of other features of the glottal source. Then, using the FPD, we describe two methods to estimate a shape parameter of the Liljencrants-Fant glottal model. By comparison with state of the art methods using Electro-Glotto-Graphic signals, we show that the one of these method outperform the compared methods.

Index Terms— glottal source, glottal model, shape parameter, phase minimization, mean squared phase.

1. INTRODUCTION

In source-filter modeling of voice production, the characterization of the speech amplitude spectrum has already received many attention (e.g. through the estimation of smooth envelopes, mel-frequency cepstral coefficients). However, less descriptions of the phase spectrum exist although this information is of great interest to model the glottal source, as it will be shown in this paper for the estimation of a shape parameter of a glottal model (e.g. the Rd parameter of the Liljencrants-Fant model [1]). Such features of the speech signal can be used for voice quality assessment as well as for voice transformation and speech synthesis. In order to estimate glottal model parameters, the methods estimating the glottal source can be used like the Iterative Adaptive Inverse Filtering method (IAIF) [2] and the minimum/maximum-phase decomposition methods (Complex Cepstrum (CC) and Zeros of the Z-Transform (ZZT) based methods [3]). In a second step, a glottal model can be fitted on the estimated glottal source by minimization of the error energy using an optimization algorithm. However, such an approach do not consider the glottal model within the step separation of the glottal source and the vocal-tract filter. These two steps are assumed independent *a priori*. In order to jointly estimate the glottal parameters with the vocal-tract filter, an AutoRegressive model with eXogenous input (ARX) can be used [4]. Recently, instead of minimizing the error energy of the additive residual (the difference between the observed signal and its model) like with ARX models, we proposed to minimize the Mean Squared Phase (MSP) of the convolutive residual (the deconvolution of the observed signal by its model) [5]. However, with MSP as presented in [5] or with ARX, the estimation of the shape parameter is dependent on the estimation of the glottal pulse position.

In this paper, after a brief description of the voice production model, we present the Function of Phase-Distortion (FPD) which characterizes the glottal source independently of: the duration of the glottal pulse, its excitation amplitude, its position as well as the position of the analysis window and the influence of a minimum-phase component of the speech signal (e.g. the vocal-tract filter). Then, using the FPD, we show that the description of an estimation method of shape parameter which has been recently presented in [6] is closely related to a new estimation method presented in this paper. Indeed, whereas the first method uses an optimization algorithm to minimize an error term computed through the FPD, the second method expresses the shape parameter in a quasi-closed form of the observed data by inversion of the FPD. Finally, the reliability of these two estimation methods will be evaluated with Electro-Glotto-Graphic (EGG) signals and compared to state of the art methods. Two examples of glottal parameter estimate conclude this paper to qualitatively assess the described methods.

2. VOICE PRODUCTION MODEL

In a short analysis window (≈ 4 periods) we assume that the voiced signal is a stationary periodic signal of fundamental frequency f_0. Therefore, we can build a discrete spectrum S_h where the h bins represent all the available h-harmonics in the Fourier transform of the windowed speech signal. The fundamental frequency f_0 is estimated using the YIN method [7] and the amplitude and phase values of this harmonic model is built by peak piking as in [8]. Using this single period representation, the voice production model of the deterministic component of the speech signal can be expressed as:

$$S_h = e^{jh\phi} \cdot G_h^{Rd} \cdot C_h \cdot jh \tag{1}$$

where $e^{jh\phi}$ is a linear-phase term which defines the position of the glottal pulse in the period, G_h^{Rd} is the Liljencrants-Fant (LF) glottal model parametrized by the Rd parameter [1], C_h is the Vocal-Tract Filter (VTF) which is assumed to be minimum-phase and jh is the term representing the radiation at the lips and nostrils level which is similar to a time derivative [9]. By division in the frequency domain, the VTF can be expressed with respect to the shape parameter Rd of the glottal model:

$$C_h^{Rd} = \mathcal{E}_-\left(\frac{S_h}{G_h^{Rd} \cdot jh}\right) \tag{2}$$

where the operator $\mathcal{E}_-(.)$ is the minimum-phase realization of its argument using the real cepstrum [10] and the linear-phase component has been discarded since $\mathcal{E}_-(.)$ is based on the amplitude spectrum only. Note that $\mathcal{E}_-(.)$ is also multiplicative (i.e. $\mathcal{E}_-(A \cdot B) = \mathcal{E}_-(A) \cdot \mathcal{E}_-(B)\ \forall |A|, |B| \geq 0$). Additionally, the lips radiation creates a zero at zero-frequency in the z-transform of the voiced signal. In equation (2), the DC of the argument of $\mathcal{E}_-(.)$ is undefined. Therefore, this DC value has to be extrapolated from the lowest harmonics. Here, we simply used $C_0^{Rd} = |C_1^{Rd}|$.

3. FUNCTION OF PHASE-DISTORTION

Based on a harmonic model, the function of phase-distortion removes the linear-phase component of an observed phase spectrum such as a nonlinear distortion remains. In order to estimate this distortion, the proposed idea is to compute the 2^{nd} order frequency derivative of the phase spectrum. Then, to retrieve a representation of the 1^{st} order derivative which is linearly related to the time delay of a given frequency (like with the group-delay), the 2^{nd} order frequency derivative is integrated once. Using a harmonic model, continuous derivatives and integration are replaced by difference and anti-difference operators. At last but not least, in order to emphasize the maximum-phase component of the spectrum, the minimum-phase contribution is removed in this measurement. Therefore, for any harmonic spectrum X_h and its minimum-phase version computed with $\mathcal{E}_-(X_h)$, the Function of Phase-Distortion (FPD) $\Phi_k(X_h)$ for the harmonic k is formalized as follows:

$$\Phi_k(X_h) = \Delta^{-1}\Delta^2\angle\left(\frac{X_h}{\mathcal{E}_-(X_h)}\right) \tag{3}$$

where Δ^2 is the 2^{nd} order phase difference operator and Δ^{-1} is the phase anti-difference operator as computed below. To avoid any problem with the wrapping of the phase in a limited range, the 2^{nd} order phase difference centered at the k^{th}-harmonic is computed in the complex plane:

$$\Delta^2\angle X_h = \angle\frac{X_{h+1}\cdot X_{h-1}}{X_h^2} \tag{4}$$

Then, using the anti-difference, the previous equation leads to:

$$\Delta^{-1}\Delta^2\angle X_k = \angle\prod_{h=1}^{k}\frac{X_{h+1}\cdot X_{h-1}}{X_h^2} \tag{5}$$

Considering a speech harmonic model S_h, the FPD $\Phi_k(S_h)$ has the following properties: It is independent of the fundamental frequency (and thus the pulse duration) since the harmonic model is a f_0-normalized representation; Using the phase difference operators, it is independent of the linear-phase component of S_h (i.e. insensitive to the position of the glottal pulse and that of the analysis window); It is independent of the minimum-phase component of S_h (e.g. the VTF) because of the deconvolution of the observed signal by its minimum-phase realization; It is independent of the gain of S_h because this latter is normalized by its minimum-phase realization ($|\mathcal{E}_-(X_h)| = |S_h|$). Consequently, the FPD is only dependent on the shape of the glottal pulse.

Figure 1 shows how the FPD of the LF glottal model varies with respect to its Rd shape parameter (note that the radiation is included in the definition of the LF model, thus LF defines $G_h^{Rd}\cdot jh$).

4. METHODS FOR GLOTTAL PARAMETER ESTIMATION

4.1. The method based on phase minimization

A first method already presented in [6] is briefly described below which estimates the Rd shape parameter of the LF model. In this method, the difference between the observed signal and its model is minimized using an optimization algorithm.

This difference is expressed in the frequency domain using the convolutive residual, the deconvolution of the signal by its model:

$$R_h^{Rd} = \frac{S_h}{G_h^{Rd}\cdot\mathcal{E}_-(S_h/G_h^{Rd}\cdot jh)\cdot jh} \tag{6}$$

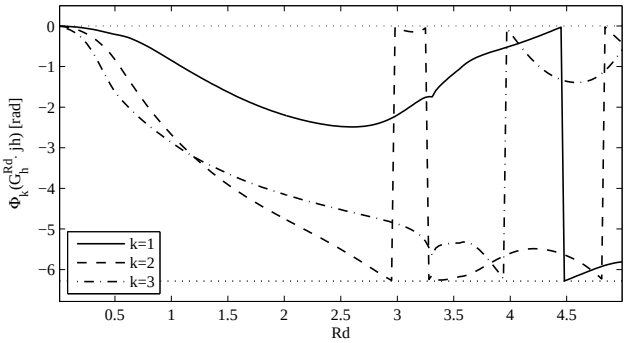

Fig. 1. The first three harmonics of the FPD of the Liljencrants-Fant model (i.e. $\Phi_k(G_h^{Rd}\cdot jh)$) with respect to the shape parameter Rd.

where the linear-phase term of equation (1) is ignored since we are not interested in the position of the glottal model and the VTF term of the voice production model has been replaced by its expression (eq. 2). The voice production model corresponds to the observed spectrum S_h if and only if the convolutive residual corresponds to a Dirac delta function. Therefore, using the phase minimization criteria [6], closer R_h^{Rd} to a unit amplitude spectrum and a zero phase spectrum, closer Rd to the optimum, the one of the observed signal. In this method, the phase spectrum of the convolutive residual is minimized through the FPD in order to estimate Rd independently of a linear-phase component. Indeed, equation (6) can be reordered in order to emphasize its similarity with the argument of the phase operator of the FPD (eq. 3):

$$R_h^{Rd} = \frac{S_h/G_h^{Rd}\cdot jh}{\mathcal{E}_-(S_h/G_h^{Rd}\cdot jh)} \tag{7}$$

since R_h^{Rd} has a unit amplitude spectrum whatever Rd, it is sufficient to minimize the phase spectrum in absolute value. Additionally, the phase difference operators, and equally the FPD, can be used to express the error of the pulse shape:

$$\Delta^{-1}\Delta^2\angle R_h^{Rd} = \Phi_k(S_h/G_h^{Rd}\cdot jh)$$

Finally, to estimate the shape parameter Rd, we minimize the following error function which is termed Mean Squared Phase using the 2^{nd} order phase Difference operator (MSPD^2):

$$\text{MSPD}^2(Rd,N) = \frac{1}{N}\sum_{k=1}^{N}\Big(\Phi_k(S_h/G_h^{Rd}\cdot jh)\Big)^2 \tag{8}$$

To obtain reliable estimate of Rd, a proper algorithm has to be chosen to minimize equation (8). Figure 2 shows MSPD^2 error functions computed on a voiced signal synthesized with a phoneme /e/ and $f_0 = 128$ Hz for three different Rd parameters. Consequently, since this error function seems to have only one minimum, a simple Brent's algorithm [11] is used to find its global minimum. The theoretical conditions of convergence of this method are discussed in [6].

4.2. The method based on FPD inversion, called FPD^{-1}

Below, a means to estimate the shape parameter of a glottal model is presented which expresses the shape parameter in a quasi closed-form of the observed data.

First, conversely to the previous method where a residual term is used, the voice production model is assumed to perfectly represent

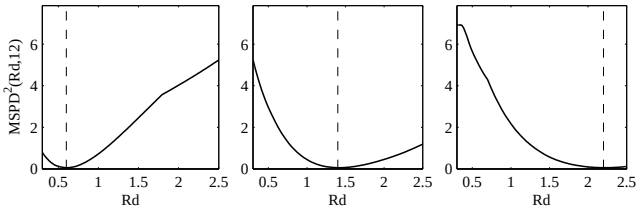

Fig. 2. Error functions $\text{MSPD}^2(Rd, 12)$ computed on synthetic signals with various optimal Rd values (dashed vertical lines).

the observed spectrum in this new method. Therefore, instead of the convolutive residual, one can write:

$$S_h = e^{jh\phi} \cdot G_h^{Rd} \cdot \mathcal{E}_- \left(\frac{S_h}{G_h^{Rd} \cdot jh} \right) \cdot jh \tag{9}$$

where the shape parameter Rd will be explicitly expressed using the other terms. Since $\mathcal{E}(.)$ is multiplicative, it can be distributed to the elements of its argument:

$$S_h = e^{jh\phi} \cdot G_h^{Rd} \cdot \frac{\mathcal{E}_-(S_h)}{\mathcal{E}_-(G_h^{Rd} \cdot jh)} \cdot jh \tag{10}$$

and therefore, one can put the observed data and the models on each side of the equality:

$$\frac{S_h}{\mathcal{E}_-(S_h)} = e^{jh\phi} \cdot \frac{G_h^{Rd} \cdot jh}{\mathcal{E}_-(G_h^{Rd} \cdot jh)} \tag{11}$$

In terms of phase-distortion, if both sides of equation (11) are equal, their respective FPD are also equal:

$$\Phi_k(S_h) = \Phi_k(G_h^{Rd} \cdot jh) \tag{12}$$

where the linear-phase term in equation (11) can be ignored since the FPD is independent of a linear-phase component. Therefore, to estimate Rd, it is sufficient to inverse $\Phi_k(G_h^{Rd} \cdot jh)$ with respect to Rd for a given harmonic k:

$$\text{given } \sigma_k = \Phi_k(S) \qquad \text{find } Rd \text{ such as } \Phi_k(G_h^{Rd} \cdot jh) = \sigma_k \tag{13}$$

However, this inversion is far from straightforward for the LF model. Indeed, the shape of the LF model is defined using intermediate parameters (see α and ϵ in [1]) which are not closed-form expressions of Rd. However, the analytic inversion can be approximated numerically. For each k-harmonic, $\Phi_k(G_h^{Rd} \cdot jh)$ can be sampled to create a lookup table whose elements are used to predict Rd from the observed σ_k value (see fig. 3 for the first harmonic). Note that according to this figure, an observed σ_k value can cross $\Phi_k(G_h^{Rd} \cdot jh)$ at multiple abscissa (e.g. 1.5 and 3.5 in fig. 3). Therefore, using only one harmonic, the shape parameter can be estimated only in an interval where $\Phi_k(G_h^{Rd} \cdot jh)$ is monotonic (e.g. $[0; 2.7]$ for $\Phi_1(G_h^{Rd} \cdot jh)$). Additionally, the interval of the observed value σ_k is limited due to the wrapping of the angle function. Therefore, a proper interval in radians for the σ_k values has to be defined where the FPD of the lookup table have the less number of discontinuities. From our observations with the LF model (see also figure 1), the interval $[0; -2\pi]$ is used in the following. Finally, a mean value can be retrieved from the Rd values predicted for each harmonic from each observed σ_k value. Algorithm 1 summarizes the whole method.

Compared to the MSPD^2 based method, a few differences exist. Firstly, it is not clear how this new approach could be used to estimate multiple parameters whereas the MSPD^2 error function can

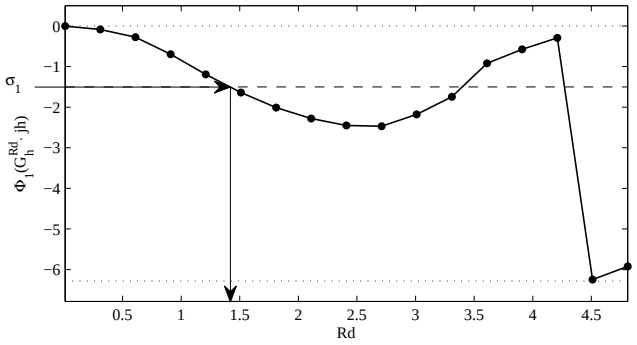

Fig. 3. Inversion of $\Phi_1(G_h^{Rd} \cdot jh)$ using a lookup table.

Algorithm 1 The method based on FPD inversion, called FPD^{-1}

Precompute the lookup table $T_k(Rd) = \Phi_k(G_h^{Rd} \cdot jh)$
for each analysis time in the speech recording **do**
  Build a harmonic model S_h on a window of 3.5 periods
  Compute $\sigma_k = \Phi_k(S)$ using eq. (3) for each harmonic $k \leq N$
  Compute Rd_k from σ_k using the lookup table $T_k(Rd)$
  Compute a mean Rd value of the estimated Rd_k values
end for

be minimized with a optimization algorithm in multiple dimensions (e.g. using preconditioned conjugate gradient or simplex). Secondly, whereas the representation of the VTF is explicit with MSPD^2 (through equations (6) and (2)), this new approach splits the VTF into two minimum-phase terms (see eq. 10). As a consequence, the DC component of the VTF is represented differently in the two methods since two extrapolations are necessary in equation (10) (see end of sec. 2). In conclusion, one can expect different results between the MSPD^2 based method and the method FPD^{-1}.

5. EVALUATION

5.1. Comparison with Electro-Glotto-Graphic signals

The Electro-Glotto-Graphy (EGG) is a non-invasive tool used in phoniatry to retrieve features of the motion of the vocal folds. Among these features, one can estimate the open-quotient O_q of the glottal pulse using the DECOM method [12]. Below, databases with synchronized waveform and EGG signals are used in order to compare the reference O_q estimated by this method and the one predicted from the estimated Rd parameter using the formula in [1]. We used four databases: APLAWD [13] which is made of 5 utterances pronounced by 10 English male and female speakers; three CMU Arctic databases [14] commonly used for speech synthesis (two male voices (*bdl* and *jmk*) and one female voice (*slt*)). Only the first 32 utterances of each Arctic database are sufficient to obtain more than 5000 comparison pairs. The APLAWD database allows to evaluate the methods among various speakers (10 speakers, 5 utterances) and the Arctic databases have a larger phonemes variation (3 speaker, 32 utterances). Six methods are compared: the one estimating jointly the position and the Rd shape parameter by minimization of the Mean Squared Phase (MSP) [5]; the two methods described in this paper (MSPD^2 and FPD^{-1}); and three methods fitting the LF model on an estimation of the glottal source by minimizing the mean squared error using a preconditioned conjugate gradient algorithm (IAIF [2], CC and the ZZT [3]) (details of the fitting procedures can

be found in [6]). The number of harmonics N taken into account in the various estimation procedures has a significant influence on the results. We computed this latter for various N values using the APLAWD database only and we selected the number of harmonics implying the smallest variance for each method: 4 for CC and ZZT, 5 for MSP, 6 for FPD^{-1}, 7 for MSPD^2 and IAIF. Figure 4 shows the corresponding results for the four databases.

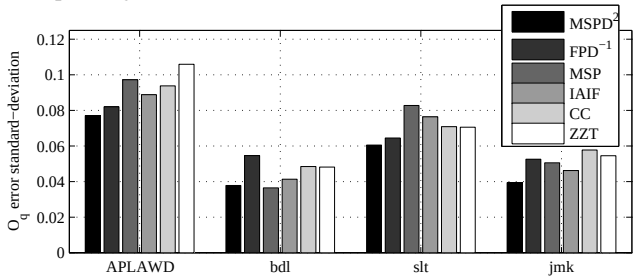

Fig. 4. Standard-deviation of the O_q estimation error.

In conclusion, whereas both methods based on MSPD^2 and FPD^{-1} outperform the four other methods on the APLAWD database, the results of the methods based on FPD^{-1} and MSP vary significantly among the voices of the Arctic databases. Conversely, the variance of the MSPD^2 based method is systematically smaller than that of all the other methods.

5.2. Examples of Rd estimates

Fig. 5 shows estimates of an open /e/. In this recording, the sound moves from a breathy phonation to a tense phonation. In addition to the Rd estimates based on MSPD^2 and FPD^{-1}, the Rd value predicted from the EGG is also shown (using formula in [1]). As expected, all these measurements move from values corresponding to lax to tense configurations.

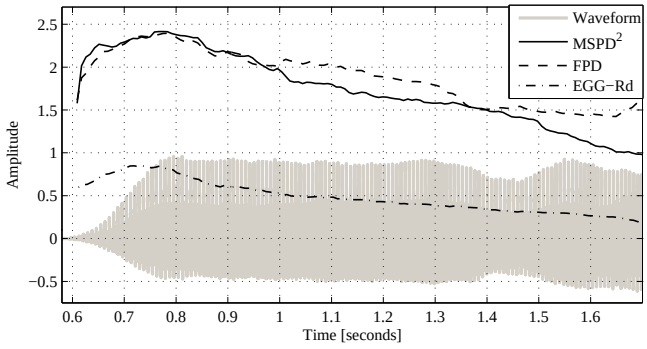

Fig. 5. Breathy to a tense phonation with Rd estimates.

Figure 6 shows the same measurements on the second utterance of the Arctic *bdl* database. Firstly, the two methods have a different systematic bias which can be explained by the differences mentioned at the end of section 4.2. Additionally, small variations of the Rd estimates are visible which are not correlated to the Rd measurement on the EGG (e.g. $t = 1.55$). Indeed, the source-filter model have some limitation, the articulatory configuration can influence the glottal source which is related to the glottal flow whereas this influence is not revealed by the EGG which is related to the vocal-folds.

6. CONCLUSIONS

In this paper, we presented the Function of Phase-Distortion (FPD) which characterizes the distortion of the phase spectrum around its linear-phase component. In the context of speech analysis, the FPD is also independent of other features of the speech signal such as it is mainly related to the shape of the pulses of the glottal source. In this paper, we used the FPD to estimate the shape parameter of a glottal model (here the Rd parameter of the Liljencrants-Fant glottal model is addressed). We showed that an estimation method which has been previously presented based on Mean Squared Phase using the 2^{nd} order phase Difference operator (MSPD^2) can be described using the FPD. Additionally, a new method based on inversion of FPD called FPD^{-1} is presented in this paper which expresses a shape parameter in a quasi-closed form of the observed data. In terms of evaluation, we used Electro-Glotto-Graphic (EGG) signals to evaluate the reliability of the described methods. By comparison with four state of the art methods, we shown that the reliability of the method FPD^{-1} varies among the used databases and the MSPD^2 based method outperforms the compared methods.

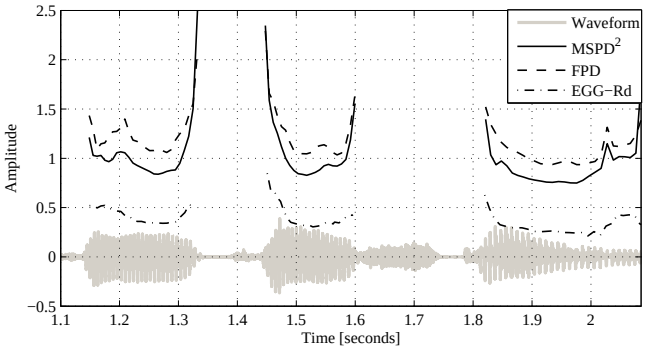

Fig. 6. Rd estimates on "particular case, Tom" from Arctic *bdl*.

7. ACKNOWLEDGMENTS

This research was partly supported by the *Affective Avatar* ANR project, by the *Respoken* FEDER project and by a grant of Centre National de la Recherche Scientifique (CNRS).

8. REFERENCES

[1] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

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

[3] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Interspeech*, 2009.

[4] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," *Interspeech*, 2005.

[5] G. Degottex, A. Roebel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *ICASSP*, 2010, pp. 5058–5061.

[6] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE ASLP*, vol. PP, no. 99, pp. 1 –1, 2010.

[7] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, April 2002.

[8] Y. Stylianou, *Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, TelecomParis, 1996.

[9] J Markel and A Gray, *Linear Prediction of Speech*, Springer Verlag, 1976.

[10] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice–Hall, 2nd edition, 1978.

[11] R. P. Brent, *Algorithms for Minimization without derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[12] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *JASA*, vol. 115, no. 3, pp. 1321–1332, 2004.

[13] G. Lindsey, A. Breen, and S. Nevard, "Spar's archivable actual-word databases," Tech. Rep., Univ. College London, U.K., 1987.

[14] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," 2003.