EVALUATION OF SPEECH INVERSE FILTERING TECHNIQUES USING A PHYSIOLOGICALLY-BASED SYNTHESIZER*

Jón Guðnason[1], Daryush D. Mehta[2, 3], Thomas F. Quatieri[3]

[1]Center for Analysis and Design of Intelligent Agents, Reykjavik University, Menntavegur 1, Iceland
[2]Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital, Boston, MA
[3]MIT Lincoln Laboratory, Lexington, MA

jg@ru.is, mehta.daryush@mgh.harvard.edu, quatieri@ll.mit.edu

ABSTRACT

Glottal inverse filtering methods are designed to derive a glottal flow waveform from a speech signal. In this paper, we evaluate and compare such methods using a speech synthesizer that simulates voice production in a physiologically-based manner that includes complexities such as nonlinear source-tract coupling. Five inverse filtering techniques are evaluated on 90 synthesized speech waveforms generated by setting six vowel configurations, three glottal models, and five fundamental frequencies. Using normalized mean square error as the primary performance metric of the estimated glottal flow derivative, results show that the accuracy of all methods depends on the configuration of the vocal tract, glottis and the fundamental frequency. Averaged over these conditions, the closed phase covariance and one weighted covariance algorithm yield lower error rates (0.41 ± 0.2) than iterative and adaptive inverse filtering (0.49 ± 0.1) and complex cepstrum decomposition (0.76 ± 0.1).

Index Terms—Glottal inverse filtering, glottal flow, glottal closure instant detection, speech signal processing, acoustics

1. INTRODUCTION

Glottal inverse filtering (GIF) is the process of deriving a glottal flow signal from acoustic and aerodynamic speech recordings [1]. This is a challenging task as it is essentially a blind source estimation problem where the input (voice source) and the system (vocal tract) are unknown. Although several promising GIF techniques have been proposed, there have been only a few reports on the comparative quantitative performance of these methods [2][3][4][5], in large part due to the challenging nature of the evaluation problem. The true glottal flow waveform (or its derivative) is rarely, if ever, measurable in practice [6], and thus quantifying the quality of a derived waveform is problematic. Indirect measures have been used, for example, by using two-channel analysis [7][8], oral flow [9] or high-speed videoendoscopy [10].

Historically, the main role of voice source–vocal tract decomposition has been in speech coding [11]; but recently, speech features obtained from the estimate of the glottal waveform have received attention more generally in the field. Voice source features have been used, for example, to improve speaker recognition [12][13] and voice transformation [14]. They have also been used to distinguish between major depressive disorders [15] and provide early diagnostic cues of Parkinson's disease [16]. Obtaining the glottal flow is also of interest in the study of voice disorders, where parameters of the glottal flow—e.g., maximum flow declination rate, minimum flow, and peak-to-peak flow—have been shown to assist clinicians in characterizing voice quality and ultimately in classifying voice disorders [17].

Motivated by the increasing importance of glottal flow estimation, the current study uses a physiologically-based speech synthesizer termed VocalTractLab[2] to evaluate five state-of-the-art GIF methods. The synthesizer produces a simulated glottal flow waveform and corresponding speech signal analogous to a microphone signal. The waveforms used in this study were formed using modal speech synthesis. The study of disordered speech remains the focus of future work. The GIF methods are then applied to the speech waveform and compared to the true glottal waveform using a normalized mean square error criterion.

2. RELATION TO PRIOR WORK

Previous studies that have used physiologically-based speech synthesis have focused on estimating parameters of the glottal flow signal [1], such as fundamental frequency [18] and formant frequencies [19]. Studies where glottal waveform estimation techniques have played a central role have typically not focused on the accuracy of the estimation technique, but rather assumed physiologically relevant features of the voice source [1][21][22]. Features include normalized amplitude quotient and closing quotient [22], the spectral difference between the first two harmonic

*This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[2] P. Birkholtz, "VocalTractLab": http://www.vocaltractlab.de

magnitudes (H1–H2), and the basic shape parameter of the Liljencrants-Fant voice source model [23].

Although synthesized speech previously has been used to obtain a quantitative comparison of glottal flow estimation techniques [2][3], the simplicity of the synthesis models applied presents a dilemma. The synthesis models typically mirrored the glottal waveform estimation techniques used in the studies. It is therefore unknown whether the techniques are simply undoing the modeled synthesis process or undoing the natural phenomena of speech production. The evaluation method presented in this paper builds on past work [4] that compares the estimated GIF waveform using reference signals generated by a physiologically-based speech synthesizer.

3. SYNTHESIS/ANALYSIS FRAMEWORK

This section describes the synthesis methods for creating the evaluation data sets, analysis methods for glottal waveform estimation, and error criterion to evaluate performance.

3.1. Synthesized data set

The study used the VocalTractLab synthesizer that is based on a 3D articulatory model of the vocal tract [24][25]. The synthesis is "bottom-up:" the glottal area and associated aerodynamics are coupled to the articulatory model, thus enabling nonlinear voice source–vocal tract coupling effects in the model outputs. The vocal tract and side cavities are modeled using a transmission line, and three types of time-domain glottal models can be selected for simulation.

Figure 1 illustrates the vowel /a/ synthesized by VocalTractLab at a sampling rate of 20 kHz. The ripple component attributed to the nonlinear source-tract coupling is observed. The shape of the vocal tract can be modified to produce different vowel sounds, and the parameters of the glottal models can simulate varying voice qualities such as modal, soft, and breathy. The glottal models have a self-oscillatory nature, and the nonlinear interaction between the vocal tract and glottis is naturally represented in the synthesis.

VocalTractLab was used to create 90 utterances for all combinations of six vowels (/a/, /e/, /ɛ/ /i/, /o/, /u/), five fundamental frequencies (f_0 = 90, 120, 150, 180, and 210 Hz), and three glottal models (Two-Mass, Geometric, and Triangular). The Two-Mass Model is the classic model, where the vocal folds are represented by two mass-spring-damper systems [26]. The Geometric Model is based on parameters that describe the shape of the glottis [27], which allows for the simulation of additional voice qualities; in this study, the Geometric Model was only set to modal (normal) voice quality. The Triangular Model is an extension of the two-mass model, where the masses are inclined as a function of the degree of abduction (hence triangular) to allow for the simulation of breathy and pressed voices [28]. In this study, the Triangular Model is only used in its normal mode.

3.2. Glottal inverse filtering analysis methods

Five state-of-the-art glottal waveform estimation techniques are compared in this paper:

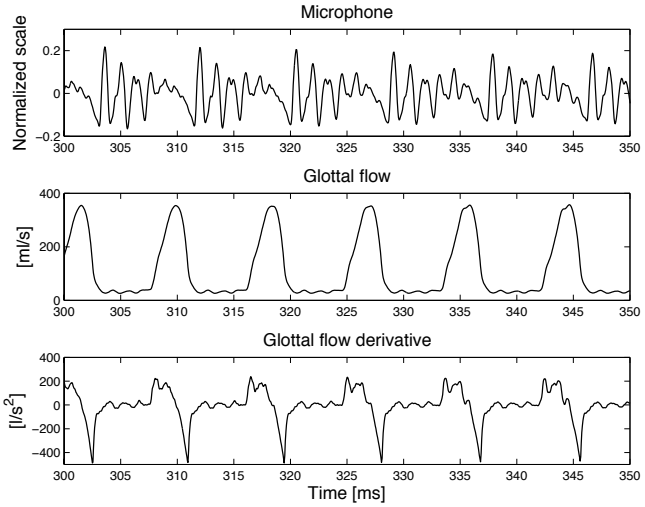

Figure 1. Exemplary waveforms from VocalTractLab generated for the vowel /a/ using the Geometric Model for the vocal folds [27].

1. Closed phase covariance analysis (CPCA) uses a hard weighting function where samples in the open phase are given zero value, and samples in the closed phase are assigned a value of one [see, e.g., [20]). The drawback of using this method is that the extent of the closed phase needs to be known through accurate identification of glottal closure instant (GCIs) and glottal opening instants (GOIs), which remains a challenging problem.

2. Weighted covariance analysis 1 (WCA1) suppresses the speech samples around the GCI using an upside-down Gaussian centered on the GCIs [29]. The method does not need the GOIs to be identified.

3. Weighted covariance analysis 2 (WCA2) also suppresses the contribution of the GCI but extends an attenuation region into the open phase. This suppresses the closing phase and the return phase around the GCI. The developers named this method, "weighted linear prediction with attenuated main excitation", [19].

4. Iterative Adaptive Inverse Filtering (IAIF) computes all-pole parameters in a few steps, each time increasing the model order, to create a successively more accurate approximation to the vocal tract transfer function and avoids over-fitting. The models are thus constrained to approximate the vocal tract without modeling the voice source [30].

5. Complex Cepstrum Decomposition (CCD) achieves a separation of the vocal tract and the voice source signal in the complex cepstrum domain by assuming that the glottis contribution is anti-causal and is therefore represented as the negative part of the quefrency domain [31].

All the methods except IAIF rely on the identification of GCIs, with CPCA also requiring the identification of GOIs.

Yet another GCI algorithm (YAGA) [32] was used to identify GCIs, and GOIs were estimated by modifying YAGA to choose the candidate nearest to the midpoint between two consecutive GCIs.

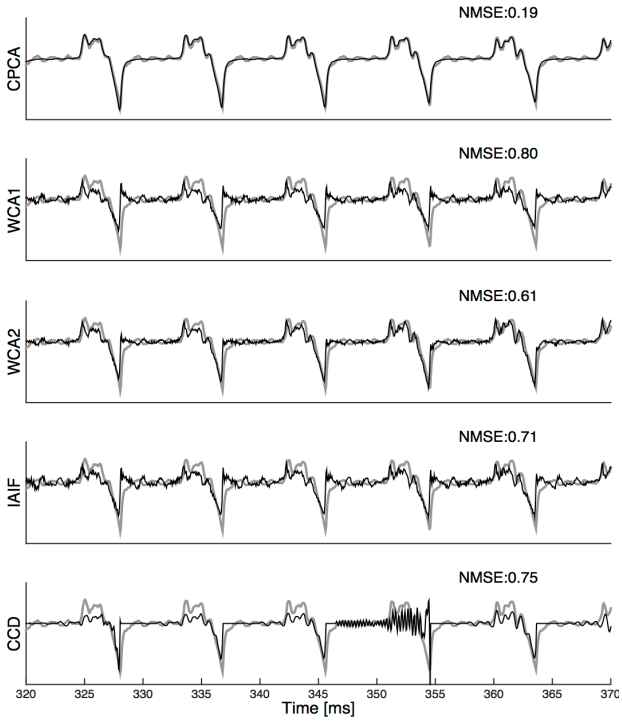

Figure 2. Illustration of glottal flow derivative estimates (black traces) plotted with the true glottal flow derivatives (gray traces) for the five GIF approaches under investigation. Normalized mean square error (NMSE) is reported for each estimate.

The first three GIF methods assessed in this paper are based on a weighted covariance analysis of speech, which obtains the all-pole vocal tract parameters $\boldsymbol{a}$ as a solution to

$$\mathbf{\Phi a} = -\xi \tag{1}$$

where the elements of the covariance matrix $\mathbf{\Phi}$ are obtained using

$$\phi_{l,k} = \sum_{n=M}^{N-1} w(n)s(n-l)s(n-k) \tag{2}$$

and the elements of the auto-covariance sequence $\boldsymbol{\xi}$ are obtained by

$$\xi_l = \sum_{n=M}^{N-1} w(n)s(n-l)s(n). \tag{3}$$

Here, *s(n)* is the speech signal, *N* is the window size in samples, *M* is the number of all-pole parameters and *l* and *k* are integers from 1 to *M*. The weighting function *w(n)* is designed to emphasize important time samples in the signal.

Figure 2 illustrates example analyses of a synthesized vowel waveform by the five GIF techniques implemented. The utterance is produced at f_0 = 120 Hz, using the vowel /a/ and the Geometric model for the glottis. The true glottal waveform derivative and its estimates using each algorithm are shown.

Table 1. Normalized mean square error (mean ± 1 standard deviation) for each of the five inverse filtering methods evaluated.

CPCA	WCA1	WCA2	IAIF	CCD
0.41 ± 0.23	0.45 ± 0.21	0.41 ± 0.14	0.49 ± 0.14	0.76 ± 0.13

3.3. Evaluation error criterion

Normalized mean square error (NMSE) was selected as an initial error criterion to provide a global metric of algorithmic performance. NMSE was defined as

$$NMSE = \frac{\sqrt{\sum_n (u(n) - G\hat{u}(n-n_d))^2}}{\sqrt{\sum_n u(n)^2}} \tag{4}$$

where *u(t)* and *û(t)* are the true and estimated glottal flow derivatives, and *n* is the time index over the stable portion of the vowel. The gain constant *G* was selected to produce the lowest NMSE. The estimated glottal flow derivative waveform was shifted by n_d samples in time to compensate for the acoustic propagation time from the glottis to the position of the synthesized microphone waveform (n_d = 14 samples for a 0.7-ms shift).

4. RESULTS

For the illustrative case of Figure 2, CPCA has the lowest NMSE value of 0.19. The estimated glottal flow derivative of CPCA gives a good fit to the opening phase, its ripple, and the return phase. The other methods also capture the ripple in the opening phase but do not follow the return phase as well. The CCD algorithm produces a high NMSE value of 0.75, explained both by consistent underestimation of the amplitude in the opening phase and a high-frequency artifact evident in the fourth glottal cycle.

Figure 3 plots the NMSE as a function of fundamental frequency for each of the five GIF methods. A general trend of decreasing performance with higher fundamental frequency is observed. Obtaining GCI and GOI is more challenging at higher frequencies, which may explain the lower performance of the methods that rely on GCI and GOI estimation. These findings are consistent with those in the literature [1][19][32]. There is also a difference in performance between methods depending on which vowels are being modeled. The IAIF method, for example, performs better on the close and near-close vowels (/u/, /o/ and /i/) than on the open and near open vowels (/a/, /e/, /ɛ/). In contrast, CPCA performs better on the open vowels than the close ones.

Figure 3 also shows the performance difference across analysis methods for three glottal models. The average NMSE over all analysis methods, vowels, and fundamental frequencies is 0.45 ± 0.14 for the Two-Mass Model, 0.45 ± 0.18 for the Triangular Model, and 0.58 ± 0.241 for the Geometric Model. GIF of waveforms synthesized with the Geometric Model thus appears to be more challenging than analysis of the other glottal models. The relative performance was maintained when average NMSE was computed within each analysis method.

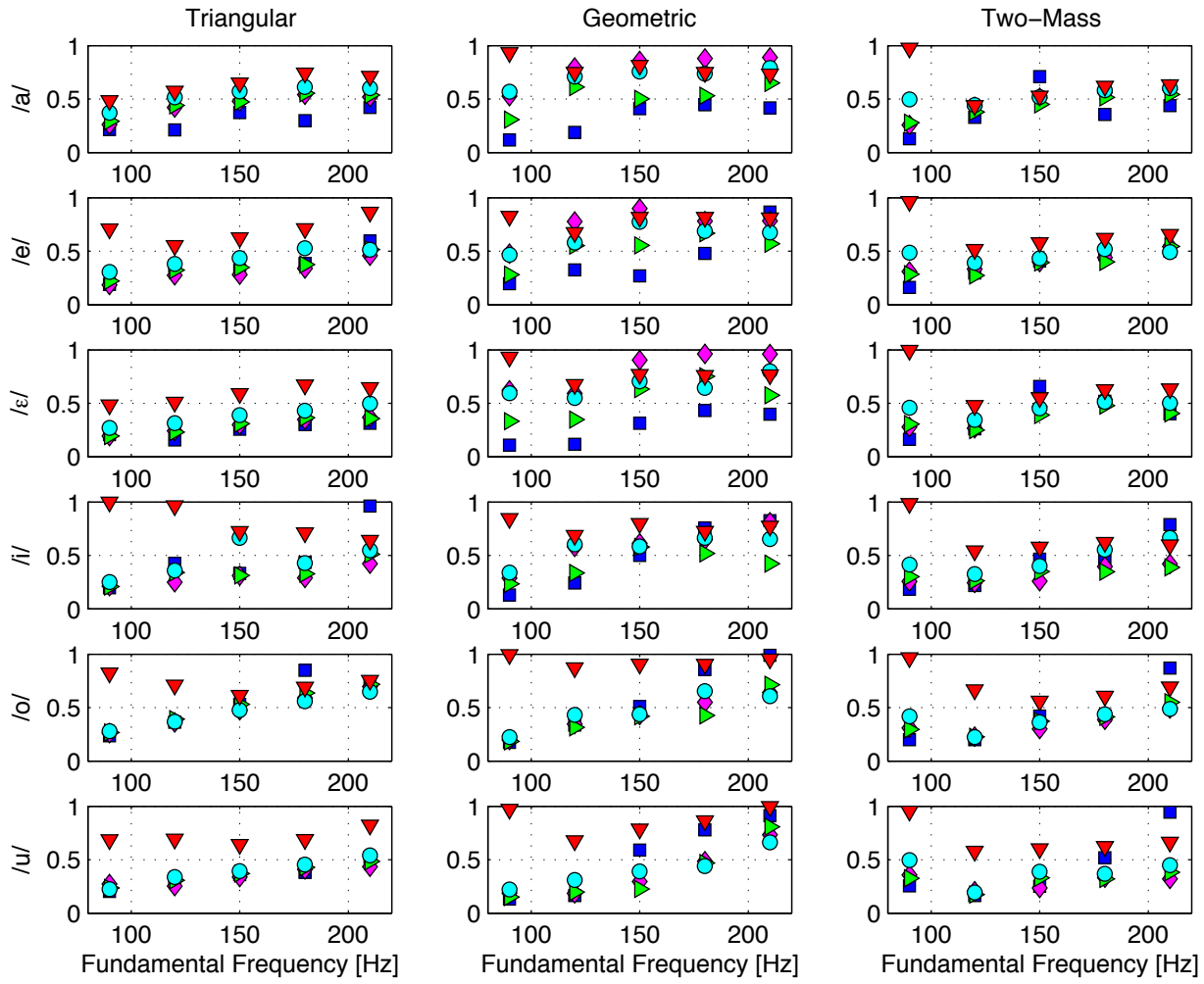

Figure 3. Normalized mean square error across five fundamental frequencies for particular synthesis configurations of six vowel types (rows) and three glottal models (columns). For each configuration, the error is plotted for the five GIF algorithms: CPCA, WCA1, WCA2, IAIF and CCD.

Table 1 shows the overall error averaged across all synthesis conditions. NMSE varied significantly depending on vowel, glottal model, and fundamental frequency, with error lowest for CPCA and WCA2 and highest for CCD.

5. CONCLUSION

Five GIF methods were assessed using the physiologically-based speech synthesizer VocalTractLab. The glottal flow derivative estimates were compared against the true glottal flow derivative waveforms produced by the synthesizer with NMSE as an initial error criterion. Voice samples were generated for six vowels, five fundamental frequencies, and three glottal models with results summarized in Fig. 3.

Increasing fundamental frequency remains a challenge for all methods of GIF. Also, utterances produced by using the Geometric glottal model appeared to be more difficult to analyze than waveforms synthesized with the other glottal models. The CPCA algorithm performed well on open vowels, whereas the IAIF algorithm performed well on closed vowels. Results also showed that the performance of all GIF methods was dependent on how the utterance was generated with respect to vowel type, glottal model, and fundamental frequency. Overall, CPCA and WCA2 were shown to perform better with respect to NMSE than the other methods, although the varying degree of performance across synthesis configurations indicates that much more work is needed for robust GIF performance.

Future research efforts warrant assessment using additional error criteria, such as standard parameters of the glottal flow waveform and its derivative (e.g., maximum flow declination rate and the coarse/fine structure of the waveform). The ability of different algorithms to estimate complementary aspects of the voice source (e.g., open phase versus closed phase properties), as well as non-modal glottal flow shapes, is also of interest.

REFERENCES

[1] P. Alku, "Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 36, no. October, pp. 623–650, 2011.

[2] N. Sturmel, C. D'Alessandro, and B. Doval, "Glottal parameters estimation on speech using the zeros of the Z-transform," *Interspeech*, pp. 665–668, 2010.

[3] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26. pp. 20–34, 2012.

[4] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatr. Logop.*, vol. 58, no. 2, pp. 102–113, 2006.

[5] D. T. W. Chu, K. Li, J. Epps, J. Smith, and J. Wolfe, "Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics.," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. EL358–62, 2013.

[6] H. Kataoka, S. Arii, Y. Ochiai, T. Suzuki, K. Hasegawa, and H. Kitano, "Analysis of human glottal velocity using hot-wire anemometry and high-speed imaging.," *Ann. Otol. Rhinol. Laryngol.*, vol. 116, no. 5, pp. 342–8, May 2007.

[7] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoust.*, vol. 33, pp. 369–377, 1985.

[8] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust.*, vol. 34, no. 4, pp. 730–743, 1986.

[9] J. Guðnason, D. D. Mehta, and T. F. Quatieri, "Closed phase estimation for inverse filtering the oral airflow waveform," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 920 – 924.

[10] Y.-L. Shue and A. Alwan, "A new voice source model based on high-speed imaging and its application to voice source estimation," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5134–5137.

[11] J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[12] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[13] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2008, pp. 4821–4824.

[14] Y. Stylianou, "Voice transformation: A survey," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2009, pp. 3585–3588.

[15] T. F. Quatieri, N. Malyska, and A. International Speech Communications, "Vocal-source biomarkers for depression: A link to psychomotor activity," *Interspeech*, pp. 1058–1061, 2012.

[16] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 1264–1271, 2012.

[17] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods.," *Curr. Opin. Otolaryngol. Head Neck Surg.*, vol. 16, pp. 211–215, 2008.

[18] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, pp. 2885–2901, 2014.

[19] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction.," *J. Acoust. Soc. Am.*, vol. 134, pp. 1295–1313, 2013.

[20] D. Y. Wong, J. D. Markel, and A. H. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust.*, vol. 27, no. 4, pp. 350–355, 1979.

[21] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive filtering," *Speech Commun.*, vol. 11, pp. 109–118, 1992.

[22] T. Backstrom, P. Alku, and E. Vilkman, "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 186–192, Mar. 2002.

[23] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *STL-QPSR*, vol. 36, 1995.

[24] P. Birkholz, "VocalTractLab." [Online]. Available: http://www.vocaltractlab.de. [Accessed: 05-Sep-2014].

[25] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS One*, vol. 8, 2013.

[26] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.

[27] I. R. Titze, "A four-parameter model of the glottis and vocal fold contact area," *Speech Commun.*, vol. 8, no. 3, pp. 191–201, Sep. 1989.

[28] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Interspeech*, 2011, pp. 2681–2684.

[29] V. Khanagha and K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, 2013.

[30] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, Jun. 1992.

[31] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, 2009, pp. 116–119.

[32] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 1, pp. 82–91, 2012.