GLOTTAL INVERSE FILTERING USING STABILISED WEIGHTED LINEAR PREDICTION

George P. Kafentzis[1], *Yannis Stylianou*[1], *and* *Paavo Alku*[2]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2]Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland
email: kafentz@csd.uoc.gr, yannis@csd.uoc.gr, and paavo.alku@tkk.fi

ABSTRACT

This paper presents and evaluates an inverse filtering technique of the speech signal which is based on the Stabilized Weighted Linear Prediction (SWLP) of speech [1]. SWLP emphasizes the speech samples that fit the underlying speech production model well, by imposing temporal weighting of the square of the residual signal. The performance of SWLP is compared to the conventional Linear Prediction based inverse filtering techniques, such as the Autocorrelation and Closed Phase Covariance method. All the inverse filtering approaches are evaluated on a database of speech signals generated by a physical model of the voice production system. Results show that the estimated glottal flows using SWLP are closer to the original glottal flow than those estimated by the Autocorrelation approach, while its performance is comparable to the Closed Phase Covariance approach.

Index Terms— Inverse filtering, Linear prediction, Closed Phase analysis, Speech analysis.

1. INTRODUCTION

Inverse filtering is a widely known method for voice and speech analysis, which mainly focuses on estimating the source of voiced speech, the glottal volume velocity waveform (or glottal airflow). The idea behind inverse filtering is to form a computational model for the vocal tract signal and then to cancel its effect from the speech waveform by filtering the speech signal through the inverse of the model. This makes apparent that inverse filtering is greatly dependent on robust vocal tract filter estimation. Inverse filtering has been extensively used in basic research of speech production and in speech synthesis, but it is awakening increasing interests also in the areas of environmental voice care of the emotional content of speech.

Since the first proposal by Miller [2], there have been several methods of inverse filtering in the literature. A number of them use additional information except from the speech signal itself, such as the electroglottographic (EGG) signal [3]. Other techniques include iterative methods to robustly estimate the glottal flow [4], while there are also approaches on joint estimation of the vocal tract system, modeled as an Autoregressive (AR) process, and the parameters of the glottal flow [5]. Most of the methods rely on Linear Prediction (LP) analysis of speech. LP is a well-known all-pole method for estimating the vocal tract signal, and there are two ways to compute it: autocorrelation and covariance method [6], but they both suffer from drawbacks. The Autocorrelation method produces a stable but biased solution for the vocal tract for a limited size window analysis. The Covariance method does not guarantee the stability of the estimated filter but it may produce an unbiased solution for limited size window analysis. Therefore, the Covariance approach is most suitable for the analysis of speech during the closed phase (i.e., when the glottis is closed) where the autoregressive hypothesis for the production of the speech signal is most valid, and this is referred to as Closed Phase Covariance method. The main issue there, is the estimation of the closed phase from the speech signal.

A fundamental problem in comparing the effectiveness of the current inverse filtering approaches and furthermore in developing new inverse filtering techniques for speech, is the fact that the true glottal air-flow signal is not known. A common approach to cope with this problem is to assess the performance of inverse filtering by using synthetic speech signals that have been created using a known, artificial waveform of the glottal excitation. However, this kind of evaluation is not truly objective because speech synthesis and inverse filtering analysis are both typically based on the source-filter model of the human voice production system.

In this paper, the performance of a recently developed all-pole method for speech recognition, referred to as Stabilized Weighted Linear Prediction, is discussed for the purpose of inverse filtering of speech and its performance is compared to the conventional LP techniques such as the Autocorrelation based and the Closed Phase Covariance based inverse filtering approaches.To overcome the aforementioned problem of knowing the true glottal airflow, experiments were conducted on a database of speech signals generated by a physical model of the vocal folds and the vocal tract suggested by Titze and Story [7]. The major advantage of this database is that both the speech pressure signal and the glottal excitation signal are available. By using the simulated speech pressure waveform as an input to an inverse filtering method, it is possible to determine how closely the obtained estimate of the voice source matches the simulated glottal flow. Time and frequency domain measures are applied on the original and the estimated glottal flows in order to quantify the similarity of the waveforms. It is shown that SWLP outperforms the conventional autocorrelation-based inverse filtering approach, while its performance is comparable to the closed phase covariance-based inverse filtering method, if in the latter case the closed phase is accurately identified from the speech signal.

The rest of this paper is organized as follows. In Section 2, the Stabilized Weighted Linear Prediction, SWLP, is quickly reviewed and its properties that make it convenient for inverse filtering are discussed. In Section 3, the inverse filtering procedure is described, whereas in Section 4, the inverse filtering performance of SWLP compared to conventional LP approaches is demonstrated. Finally, Section 5 concludes the paper.

This work was partially supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623

2. OVERVIEW OF SWLP

Stabilized Weighted Linear Prediction (SWLP) was introduced by Magi et al. [1], as an all-pole modeling method based on the Weighted Linear Prediction (WLP) [8]. A quick review of WLP and SWLP is following next.

2.1. Weighted Linear Prediction, WLP

As in conventional LP, sample $x[n]$ is estimated by a linear combination of the past p samples:

$$\hat{x}[n] = -\sum_{i=1}^{p} a_i x[n-i], \tag{1}$$

where the coefficients $a_i \in \Re$. The prediction error $e_n(\mathbf{a})$, or the residual, is defined as

$$e_n(\mathbf{a}) = x[n] - \hat{x}[n] = x[n] + \sum_{i=1}^{p} a_i x[n-i] = \mathbf{a}^T \mathbf{x}[n], \tag{2}$$

where $\mathbf{a} = [a_0\, a_1 \cdots a_p]^T$ with $a_0 = 1$ and $\mathbf{x}[n] = [x[n] \cdots x[n-p]]^T$.

The prediction error energy $E(\mathbf{a})$ in the WLP method is given by

$$E(\mathbf{a}) = \sum_{n=1}^{N+p} (e_n(\mathbf{a}))^2 w_n = \mathbf{a}^T \Big(\sum_{n=1}^{N+p} w_n \mathbf{x}[n]\mathbf{x}^T[n] \Big) \mathbf{a} = \mathbf{a}^T \mathbf{R} \mathbf{a}, \tag{3}$$

where w_n is the weight imposed on sample n, N is the length of the signal $x[n]$, and $\mathbf{R} = \sum_{n=1}^{N+p} w_n \mathbf{x}[n]\mathbf{x}^T[n]$. This is a constrained minimization problem:

$$\text{minimize } E(\mathbf{a}) \text{ subject to } \mathbf{a}^T \mathbf{u} = 1,$$

where $\mathbf{u}$ is the vector defined as $\mathbf{u} = [1\ 0 \ldots 0]^T$. It can be seen that the autocorrelation matrix $\mathbf{R}$ is weighted, in contrast to the conventional LP analysis. Because of this weighting function, matrix $\mathbf{R}$ in (3) is symmetric but not Toeplitz. However, it is positive definite, and this makes the minimization problem convex. Using Lagrange multipliers, it can be shown that $\mathbf{a}$ satisfies the linear equation

$$\mathbf{R}\mathbf{a} = \sigma^2 \mathbf{u}, \tag{4}$$

where $\sigma^2 = \mathbf{a}^T \mathbf{R} \mathbf{a}$ is the error energy. Finally, the WLP all-pole model is obtained as $H(z) = 1/A(z)$, where $A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$.

2.2. Weighting function

The time domain weighting function w_n is the key point of WLP. In [8], the weighting function was chosen to be the Short-Time Energy (STE)

$$w_n = \sum_{i=0}^{M-1} x[n-i-1]^2, \tag{5}$$

where M is the length of the STE window. The use of the STE window can be justified as following. STE emphasizes the speech samples of large amplitude which typically occur during the closed phase interval of glottis. It is well-known that applying LP analysis on speech samples that belong to the glottal closed phase interval will generally result in a more robust spectral representation of the vocal tract since the hypothesis that speech samples have been produced by an autoregressive process is more valid. Therefore, increasing the weight on these samples that occur during the glottal closed phase it is more likely to estimate accurately the vocal tract filer. This makes WLP a promising method for inverse filtering. Note that closed phase (CP) covariance techniques also exploit the CP interval. However, they typically suffer from lack of robustness in the identification of the CP interval. In Fig.1, the correlation between the glottal closed phase and the STE weighting function is illustrated on a clean vowel.

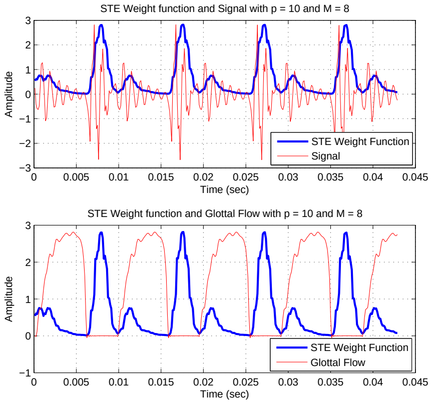

Fig. 1. Upper panel: time domain waveforms of speech (vowel /a/ produced by male speaker) and short-time energy (STE) weight function (M=8). Lower panel: Glottal flow waveform of the vowel /a/ together with the STE weight function (M=8).

2.3. Stabilized WLP, SWLP

The WLP method with the STE window does not ensure stability of the all-pole model. Therefore, in [1], a formula for a generalized weighting function to be used in WLP is developed in order to guarantee stability. The autocorrelation matrix $\mathbf{R}$ in (4) can be expressed as

$$\mathbf{R} = \mathbf{Y}^T \mathbf{Y}, \tag{6}$$

where

$$\mathbf{Y} = [\mathbf{y_0}\, \mathbf{y_1} \cdots \mathbf{y_p}] \in \Re^{(N+p)x(p+1)}$$

and

$$\mathbf{y}_0 = [\sqrt{w_1}x[1] \cdots \sqrt{w_N}x[N]\ \ 0 \cdots 0]^T.$$

The column vectors are given by

$$\mathbf{y}_{k+1} = \mathbf{B}\mathbf{y}_k,\ k = 0, 1, \cdots, p-1, \tag{7}$$

where $\mathbf{B}$ is a matrix with all elements zero except the secondary diagonal of the matrix which defined for all $i = 1, \cdots, N+p-1$ as

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \le w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases}$$

The WLP method using matrix $\mathbf{B}$ is referred to as *Stabilized Weighted Linear Prediction*, and it can be shown that the obtained all-pole filter is always stable [1].

3. INVERSE FILTERING PROCEDURE AND EVALUATION MEASURES

All the inverse filtering approaches studied here were applied on a database of sustained vowels generated by a physical model of the human voice production system [7]. By using the simulated speech pressure waveform as an input to an inverse filtering method, it is possible to determine how closely the obtained estimate of the voice source matches the simulated glottal flow. The sound pressure and glottal flow waveforms were generated with a computational model of the vocal folds and acoustic wave propagation and were digitized using a sampling frequency of $F_s = 8kHz$ and precision of 16bits. In detail, self-sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements. The input parameters to the model consisted of lung pressure, prephonatory glottal half-width (adduction), resting vocal fold length and thickness, and normalized activation levels of the crycothyroid (CT) and thyroarytenoid (TA) muscles, which were then transformed into mechanical parameters for the model, according to [7]. Both adult male and female speech were produced produced by modifying the resting vocal fold length and activation levels of the CT and TA muscles. Eight different fundamental frequency values ($105, 115, 130$, and 145 Hz for adult male speech, and $205, 210, 230$ and 255 Hz for female speech) for each vowel were generated.

The order for the autoregressive process for all the considered inverse filter approaches was $p = 10$. Specifically we consider two standard approaches: the autocorrelation method, which will be noted here as *LPC*, and the Closed Phase Covariance method, noted as *CovLPC*. For the suggested SWLP approach, the parameter M was set to 8 and 24, a relatively low and high value for M (according to previous works on SWLP, i.e., [1]), in order to investigate the role of M during the inverse filtering process. The different choices for M will be noted as $SWLP_8$ and $SWLP_{24}$, for $M = 8$ and $M = 24$, respectively.

The analysis window was set to 250 ms for the autocorrelation based approaches (i.e., $SWLP_8$, $SWLP_{24}$, and LPC). while for CovLPC this was determined by the detected closed phase interval. The closed phase interval was determined by the stability of the first formant, regarding its frequency, as suggested in [9]. A typical Hanning window was used for the autocorrelation based approaches while a rectangular window was used for the CovLPC method. The frame rate of the analysis was set equal to one pitch period for all methods. Before estimating the all-pole filter, the lip radiation effect was canceled by a first order all-pole filter with its pole at $z = 0.999$. Specifically for CovLPC and because the covariance method does not guarantee stability of the estimated all-pole filter, the poles of the estimated filter were computed and those which were located outside the unit circle were simply replaced by their corresponding mirror image inside the unit circle, while the poles on the positive real axis were removed. This modified filter was then used for inverse filtering.

Using the estimated filters (i.e., $SWLP_8$, $SWLP_{24}$, LPC, CovLPC), the inverse filtered speech signals were computed in a frame by frame basis. For this, a window of two local pitch periods and a frame rate of one pitch period was applied. The overall glottal flow was synthesized using the Overlap-Add (OLA) method.

The selected evaluation measures for assessing the performance of the inverse filtering techniques were the Signal to Reconstruction Error ratio (SRER), and H1-H2. SRER is a standard index for measuring the effectiveness of modeling a waveform and is defined as:

$$\mathrm{SRER} = 20\log_{10}\left(\frac{\sigma_{s[n]}}{\sigma_{e[n]}}\right) \tag{8}$$

where $s[n]$ is the original (or true) glottal flow signal in our case, $e[n]$ is the modeling (or reconstruction) error, $e[n] = s[n] - \hat{s}[n]$, and σ denotes the corresponding standard deviation. SRER was computed from the overall glottal flow waveforms.

H1-H2 is a frequency domain metric and is derived from the spectrum of the glottal flow [10]. H1-H2 is defined as the difference in decibel between the amplitudes of the fundamental and the second harmonic of the source spectrum. H1-H2 is an index of the spectral decay (or spectral tilt) of the glottal spectrum. To compare the estimated glottal airflow with the true glottal airflow using the above frequency domain metric, we suggest to measure the difference between these two measurements from the estimated and the true glottal airflow. More specifically we suggest:

$$ER_{H1H2} = \left|Ref_{H1H2} - Est_{H1H2}\right| \tag{9}$$

where Ref_{H1H2} and Est_{H1H2} denote the H1-H2 metric for the true (or reference) and the estimated glottal airflow, respectively. For a good estimation ER_{H1H2} should be close to zero.

These specific measures were selected for the evaluation of the inverse filtering approaches because they can be automatically extracted from the data without any user adjustment.

4. RESULTS

In our experiments, 4 different vowels, /aa/, /ae/, /eh/, and /ih/, with 8 different fundamental frequencies for each vowel were used: $105, 115, 130, 145, 205, 210, 230$, and 255 Hz. Two characteristic examples of the inverse filtered waveforms for two fundamental frequencies, 105 and 230 Hz, of the vowels /aa/ and /eh/ are shown in Fig.2 and Fig.3, respectively. Based on the results shown in Fig.2 and Fig.3, the glottal flow estimate based on SWLP with $M = 24$ is closer to the original glottal flow than that of the conventional autocorrelation method (LPC) that estimate is very close to CovLPC glottal flow estimate. This is true for all vowels and f_0s in our experiments, for frames where the closed phase interval is accurately identified. For low pitch vowels (as in Fig.2 with $f_0 = 105$ Hz), SWLP with $M = 24$ glottal flow estimate shows a decreased ripple in the closed phase interval than the conventional autocorrelation approach (LPC). For higher pitch vowels (as in Fig.3, for $f_0 = 230$ Hz), the closed phase interval is smaller, and thus the samples with high amplitude that typically belong to that interval are less than in the lower pitch cases. In these cases, a low value of M is more suitable. However, a low value of the M parameter influences in a

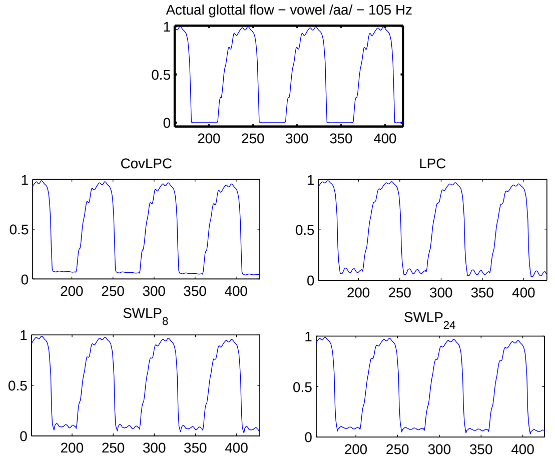

Fig. 2. Glottal flow estimates for vowel /aa/ of f_0 = 105 Hz. Upper panel: Original glottal flow. Middle panel: Covariance (left) and Autocorrelation (right) based glottal flow estimates. Lower panel: SWLP with $M = 8$ and $M = 24$ glottal flow estimates. In all panels, time is indicated in samples.

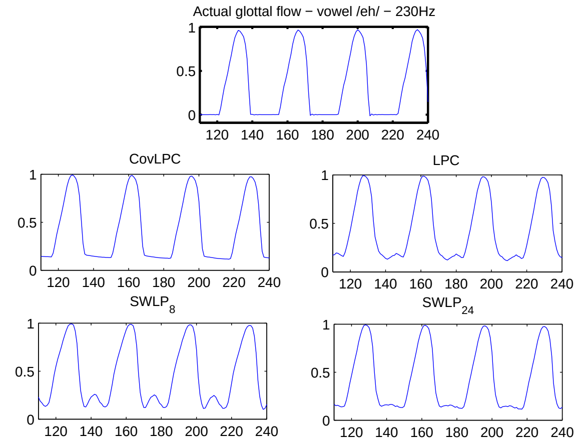

Fig. 3. Glottal flow estimates for vowel /eh/ of f_0 = 230 Hz. Upper panel: Original glottal flow. Middle panel: Covariance (left) and Autocorrelation (right) based glottal flow estimates. Lower panel: SWLP with $M = 8$ and $M = 24$ glottal flow estimates. In all panels, time is indicated in samples.

negative way the estimation of the vocal tract filter as it was shown in [1]. Indeed, the behavior of SWLP in spectral modeling depends on the M parameter: for lower M values, SWLP shows a smooth spectral behavior, whereas for higher M values, the sharpness of the resonances in the SWLP spectrum increases. Therefore, for high pitch speakers, it would be interesting to investigate the combination of contiguous closed phase speech samples in order to be able to increase the value for M.

To compare the estimated glottal air-flow signals with the original ones, the aforementioned measures for all vowels and frequencies were used and their mean and standard deviation values are shown in Tables 1 and 2, for SRER and ER_{H1H2}, respectively. Based on the results listed in Table 1, $SWLP_{24}$ provides better results than $SWLP_8$ in all cases. It also provides better results than LPC except in the case of $/ae/$. For this time domain criterion, the Closed Phase Covariance approach (CovLPC) provides the best results with higher, however, standard deviation, following from the suggested approach. Note that in this case, however, the closed phase interval was estimated and used explicitly by CovLPC while $SWLP_{24}$ was using that information only implicitly (through the weighting function). Using the frequency domain criterion, Table 2 shows that $SWLP_{24}$ outperforms significantly the LPC approach while it is quite close to the performance provided by CovLPC. However, note also in this case the high standard deviation for CovLPC in contrast to the low standard deviation obtained for $SWLP_{24}$.

	SRER			
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC
$/aa/$	33.5 (±2.0)	39.7 (±4.5)	36.2 (±5.7)	41.9 (±6.3)
$/ae/$	32.7 (±4.4)	35.2 (±2.9)	37.8 (±3.0)	40.4 (±6.4)
$/eh/$	34.0 (±1.9)	38.4 (±4.2)	33.9 (±4.0)	40.5 (±5.2)
$/ih/$	32.3 (±1.5)	37.6 (±3.1)	35.3 (±4.6)	39.2 (±5.6)

Table 1. Mean and standard deviation of the SRER value for each vowel (all 8 frequencies) and method is illustrated.

	ER_{H1H2}			
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC
$/aa/$	0.68 (±0.10)	0.23 (±0.09)	0.75 (±0.09)	0.20 (±0.20)
$/ae/$	0.15 (±0.12)	0.15 (±0.05)	0.55 (±0.05)	0.18 (±0.13)
$/eh/$	0.34 (±0.09)	0.30 (±0.07)	0.54 (±0.08)	0.38 (±0.17)
$/ih/$	0.72 (±0.14)	0.39 (±0.11)	0.85 (±0.12)	0.35 (±0.24)

Table 2. Mean and the standard deviation of ER_{H1H2} for each vowel (all 8 frequencies) and each method is illustrated.

5. CONCLUSIONS

In this paper, we discussed the performance of the Stabilized Weighted Linear Prediction in inverse filtering speech waveforms. This method applies temporal weighting on the square of the residual signal, and thus emphasizing the samples of high energy, which typically belong to the closed phase interval during phonation. The method was tested on a database produced by physical modeling of the voice production system. Using a time domain and a frequency domain criterion, it was shown that the glottal flow waveforms obtained by SWLP are closer to the original glottal flow waveform than those obtained by the conventional autocorrelation linear prediction method and is comparable to the conventional closed phase covariance method.

Acknowledgements

The authors would like to thank Prof. Brad Story for providing them the database of physically modeled speech waveforms.

6. REFERENCES

[1] C. Magi J. Pohjalainen T. Backstrom and P. Alku. Stabilised Weighted Linear Prediction. *Speech Communication*, 51:401–411, 2009.

[2] R. Miller. Nature of the Vocal Cord Wave. *J. Acoust. Soc. Am.*, 31:667–677, 1959.

[3] D. Veeneman and S. BeMent. Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals. *IEEE Trans. Acoust. Speech Signal Process.*, 33:369–377, 1985.

[4] P. Alku. Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*, 11:109–118, 1992.

[5] P. Milenkovic. Inverse Filtering by Joint Estimation of an ar System with a Linear Input Model. *IEEE Trans. Acoust. Speech Signal Process.*, 34:28–42, 1986.

[6] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, Engewood Cliffs, NJ, 2002.

[7] I. Titze and B. Story. Rules for Controlling Low-dimensional Vocal Fold Models with Muscle Activites. *J. Acoust. Soc. Am.*, 112:1064–1076, 2002.

[8] C. Ma Y. Kamp and L.Willems. Robust Signal Selection for Linear Prediction Analysis of Voiced Speech. *Speech Communication*, 12:69–81, 1993.

[9] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–587, 1999.

[10] I. Titze and J. Sundberg. Vocal Intensity in Speakers and Singers. *J. Acoust. Soc. Am.*, 107:581–588, 1992.