

ALGORITHM FOR AUTOMATIC GLOTTAL WAVEFORM ESTIMATION WITHOUT THE RELIANCE ON PRECISE GLOTTAL CLOSURE INFORMATION

Elliot Moore, Mark Clements

Georgia Institute of Technology
Department of Electrical and Computer Engineering
Atlanta, GA 30032

ABSTRACT

An automated glottal waveform estimation algorithm is presented that improves on a previous manual glottal extraction technique which produced excellent glottal waveform estimates. The algorithm uses only basic approximations of glottal closure regions and successive iterations to find the best candidate for a glottal waveform estimate within a speech frame. Visual comparisons of the glottal waveform estimates created by the algorithm and those generated from the use of glottal closure information provided by an electroglottograph (EGG) reveal that the algorithm produced virtually identical estimates.

1. INTRODUCTION

One of the more complex processes in speech analysis is the estimation of the glottal waveform. The primary problem in estimating the glottal waveform lies in the difficulty of separating glottal and vocal tract characteristics in the acoustic speech waveform. Performing closed-phase analysis implies estimating vocal tract characteristics from a region of the speech signal where the vocal folds are assumed closed and the interaction of the glottal and vocal tract dynamics are reduced. Essentially, this requires that the glottal closure instants (GCI) be identified directly from the acoustic speech signal. Various studies have investigated identifying glottal closure instants (GCI) using techniques based on dynamic programming [1], glottal input power [2], formant stability [3], and residual energy [4]. However, identifying glottal closure instants directly from the acoustic waveform is complicated by numerous factors including speaking style and gender. Females tend to exhibit a higher pitch than males requiring a more rapid motion of the glottis which does not always yield complete closure. Vocal disorders and emotional stress can affect the accuracy of identifying specific instants of glottal closure, assuming that one exists. External sensors, such as Electroglottographs (EGG), have been documented to correlate well with the mechanics of glottal motion and therefore yield fairly accurate estimates of glottal closure. However, it is necessary to collect

data from these sensors concurrently with the acoustic data which is not desirable.

Due to the complexity of finding glottal closure instants (GCI) for estimating the vocal tract, research in [5] proposed a manual glottal extraction technique that produced smooth and reliable estimates of the glottal waveform over a wide range of speaking styles. The technique involved manually sliding small windows across disjoint areas of estimated glottal closure in the speech signal in order to find regions that would yield the best glottal waveform approximations. The process produced excellent glottal waveform estimates, but was hampered by the necessity of manual intervention. The algorithm in this paper borrows from the principles used in [5] but provides improvements that enable the process to be automated. The algorithm has been successfully tested and implemented for work in [6] and shown to be effective in producing smooth glottal estimates even for singing voice.

2. THEORY

A useful model of speech production consists of a cascade of linearly separable filters according to equation 1

$$S(z) = G(z)V(z)R(z) \quad (1)$$

where $S(z)$ represents the acoustic speech waveform (for voiced speech), $G(z)$ represents glottal waveform shaping of the vocal folds, $V(z)$ models the vocal tract configuration, and $R(z)$ represents the radiation at the lips. Given this model assumption, an approximation of the quasi-periodic glottal waveform ($G(z)$) could be estimated from the acoustic speech waveform if the effects of the vocal tract ($V(z)$) and lip radiation ($R(z)$) were “removed”. The effects of lip radiation can be modelled with a first order zero ($0.95 \leq z_0 < 1$) in discrete-time. Closed phase analysis is often used to estimate the vocal tract. Fig. 1 shows that the ideal glottal flow consists of an open phase (i.e., maximum interaction with vocal tract) and a closed phase (i.e., minimal interaction with vocal tract). Obtaining an accurate estimate

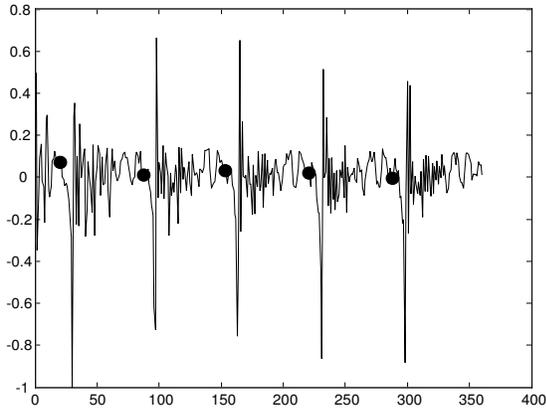


Fig. 3. LP Residual with algorithm starting points

glottal derivative estimate in \mathbf{G} . The reason for this step was based on the understanding that the first term of the LP analysis (a_1) represented the ratio of the autocorrelation at lag 1 to the autocorrelation at lag 0 according to equation 3.

$$a_1 = \frac{-r(1)}{r(0)} \quad (3)$$

In essence, a_1 represented to a certain extent how well two consecutive samples were correlated to one another. Utilizing this principle, smoother (i.e., less noisy) estimates would exhibit values of a_1 closer to 1 than an estimate that was more noisy. The coefficients of a_1 for each glottal estimate were stored in a vector (a_1). The indices (e) of the vector a_1 that corresponded to the values closest to 1 were chosen as the best glottal derivative estimates from \mathbf{G} . As can be seen for example in Fig. 4, values of a_1 that are closer to 1 are smoother in appearance than those who are not. This is particularly noticeable for values close to zero such as for $|a_1| = .056981$ and $|a_1| = 0.16141$. The indices (e) representing the top 99th percentile (i.e., the values closest to one and greater than 99% of the other choices) were used to average the best estimates of the glottal derivative waveform ($G[e]$) and LP coefficients ($A[e]$). The glottal waveform estimate was obtained by integrating the resulting glottal derivative estimate.

4. RESULTS

For the purpose of this paper, a subset of speech recordings ($f_s = 8kHz$) with the corresponding EGG data was chosen to show glottal waveform estimates generated using this algorithm compared to estimates created by using GIF with the exact glottal closure information provided by the EGG. A total of 4 subjects (2 males, 2 females) were used and a sample of the glottal waveform estimates can be seen in Figs. 5, 6, 7, and 8. In all the figures, part *a.* represents the

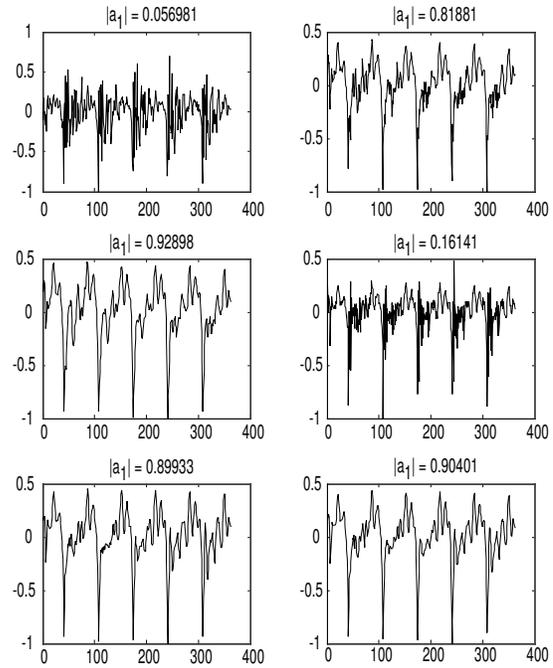


Fig. 4. Glottal derivative approximations by iteration

glottal waveform estimate from the algorithm with an LP order of $P = 10$ (i.e., 20 iterations) and part *b.* represents the estimate obtained utilizing EGG data. In nearly every case, it was observed that the estimates created by the algorithm were virtually identical to the estimates generated with the precise EGG glottal closure information.

5. CONCLUSION

This algorithm represents an improvement to the one in [5] by virtue of providing a means of unsupervised implementa-

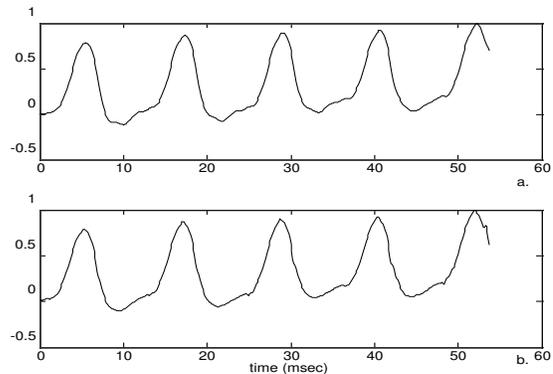


Fig. 5. Glottal Waveform Estimates for male 1: a.) Algorithm Method b.) EGG Method

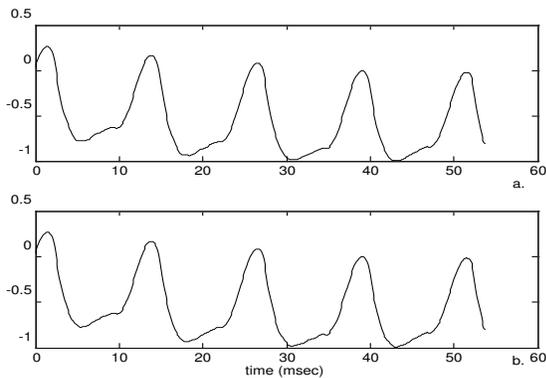


Fig. 6. Glottal Waveform Estimates for male 2: a.) Algorithm Method b.) EGG Method

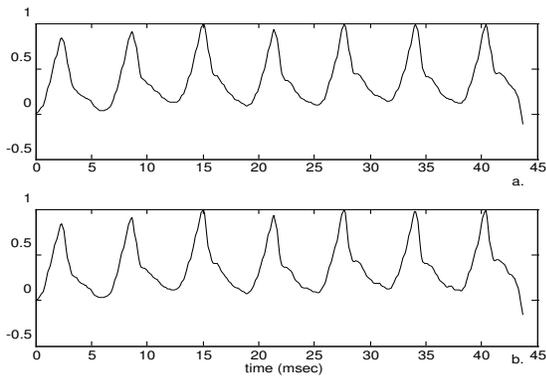


Fig. 7. Glottal Waveform Estimates for female 1: a.) Algorithm Method b.) EGG Method

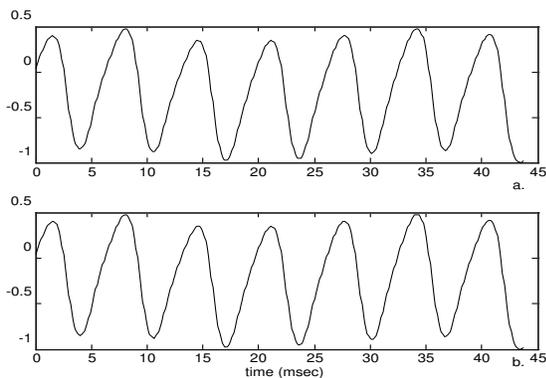


Fig. 8. Glottal Waveform Estimates for female 2: a.) Algorithm Method b.) EGG Method

tion. Essentially, the algorithm follows the same procedures that were accomplished manually in [5] with the improvement of automating the decision for the best glottal waveform estimate. In addition, this algorithm produces excellent glottal waveform estimates without the added complexity of finding precise GCI's for analysis. No assumptions about GCI's are made beyond the assumption that they are in the area of the negative peaks of the glottal waveform derivative. Therefore, the algorithm is free to search for the best "candidate" within each speech frame. This is an important point since the result produced by the algorithm is almost always the best possible approximation of the glottal waveform for a given speech frame. While it was once thought that the results produced by the algorithm were due to choosing the same estimation points as those identified by the EGG data, experiments have shown this to not strictly be the case as at times the algorithm will select points on the closing phase and sometimes on the early stages of the opening phase to find the smoothest estimate. The results also suggest that while using multiple *approximations* of glottal closure regions were necessary, finding *precise* glottal closure instants was not.

6. REFERENCES

- [1] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dyspa algorithm for estimation of glottal closure instants in voiced speech," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 2002, vol. 1, pp. 349–352.
- [2] D. M. Brookes and H. P. Loke, "Modeling energy flow in the vocal tract with applications to glottal closure and opening detection," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 1999, vol. 1, pp. 213–216.
- [3] 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 Processing*, vol. 7, no. 5, pp. 569–585, Sept 1999.
- [4] D. Y. Wong, J. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 4, pp. 350–355, August 1979.
- [5] K. E. Cummings and M. A. Clements, "Glottal models for digital speech processing: A historical review and new results," *Digital Signal Processing: A Review Journal*, vol. 5, no. 1, pp. 21–42, Jan 1995.
- [6] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proceedings, 25th Annual Conference on Engineering in Medicine and Biology*, 2003, pp. 2849–2852.