CLOSED PHASE ESTIMATION FOR INVERSE FILTERING THE ORAL AIRFLOW WAVEFORM*

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

[1]Center for Analysis and Design of Intelligent Agents, Reykyavik 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, daryush.mehta@alum.mit.edu, quatieri@ll.mit.edu

ABSTRACT

Glottal closed phase estimation during speech production is critical to inverse filtering and, although addressed for radiated acoustic pressure analysis, must be better understood for the analysis of the oral airflow volume velocity signal that provides important properties of healthy and disordered voices. This paper compares the estimation of the closed phase from the acoustic speech signal and the oral airflow waveform recorded using a pneumotachograph mask. Results are presented for ten adult speakers with normal voices who sustained a set of vowels at a comfortable pitch and loudness. With electroglottography as reference, the identification rate and accuracy of glottal closure instants for the oral airflow are 96.8 % and 0.28 ms, whereas these metrics are 99.4 % and 0.10 ms for the acoustic signal. We conclude that glottal closure detection is adequate for close phase inverse filtering but that improvements to detection of glottal opening instants on the oral airflow signal are warranted.

***Index Terms*—**Inverse filtering, glottal airflow, Rothenberg mask, glottal closure instant detection

1. INTRODUCTION

Closed-phase covariance analysis is known to provide robust estimates of the all-pole vocal tract transfer function to enable inverse filtering of acoustic speech signals [1, 2] and oral airflow [3] to extract the voice source signal. Reliable identification of glottal closure and opening instants, and thus the closed phase, from the acoustic speech signal [4, 5] has made closed-phase covariance analysis practicable without the use of an electroglottographic (EGG) signal, and avoids manual or iterative methods to estimate the closed phase of the glottal airflow [6]. It is not known, however, whether detection algorithms developed to analyze acoustic signals can be applied directly to aerodynamic recordings of airflow. With this motivation, the goal of the current work is to estimate the closed phase timing by applying acoustics-derived methods of glottal closure and opening detection to oral airflow, which is unaltered by the acoustic radiation characteristic. The ultimate goal is to attain accurate inverse filtering and yield voice source features for the analysis of healthy and disordered voices.

1.1. Motivation

Voice source features extracted from the acoustic signal have played particular roles as biomarkers for neurological disorders, including major depressive disorder [7] and Parkinson's disease [8]. Recording oral airflow during phonation using a pneumotachograph mask [9] is widely considered the most direct manner in which to estimate the volume velocity airflow exiting the mouth. Clinical voice assessment typically includes a measure of average airflow and subglottal pressure to enable the estimation of a vocal efficiency ratio relating the voice source input to the speech system output. In addition, tracking the high-bandwidth airflow waveform provides for inverse filtering methods to yield the voice source waveform and features such as maximum airflow declination rate, minimum airflow, and peak-to-peak amplitudes that are commonly linked with hyperfunctional vocal behavior [10] and other disorders.

Currently, there are several techniques for inverse filtering the acoustic voice signal recorded with a microphone in the free field that consist of linear prediction coding (LPC)–based estimates of coefficients that describe the supraglottal tract transfer function [2]. Covariance methods of linear prediction require knowledge of the instants of glottal closure and opening such that LPC coefficients are derived during the closed phase of vocal fold vibration during which the supraglottal tract is closest to having all-pole qualities. The performance of inverse filtering algorithms has been evaluated primarily on databases consisting of acoustic microphone signals and a reference signal (e.g., from an EGG or laryngograph).

* The Lincoln Laboratory portion of 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. Also supported by the National Institutes of Health National Institute on Deafness and Other Communication Disorders under grant R33 DC011588.

1.2. Relation to prior work

Closed-phase covariance analysis of speech requires that the glottal closure instant (GCI) and glottal open instant (GOI) are accurately identified for each glottal cycle in the speech signal. This is often done indirectly using a two channel analysis where the EGG signal is used to extract instants [11, 12]. A limitation with this approach is that the acoustic delay between the EGG and the signal being analyzed (microphone or oral airflow) is not known exactly. Reliable identification of the GCIs and the GOIs in the signals that are being analyzed is therefore very desirable. Closed-phase covariance analysis has also been performed without knowing the GCIs and GOIs beforehand by sliding the analysis window sample by sample and choosing the minimum energy of the resulting residual [13, 14], by minimizing formant ripple [15], or by detecting statistically significant changes in the AR coefficients [16].

One-channel analysis can be applied where algorithms such as YAGA [4] or DYPSA [17] can provide closed phase timing information. As alluded to earlier, however, it is unknown whether algorithms developed to analyze acoustic signals can be applied directly to aerodynamic recordings of airflow. If successful, single-channel inverse filtering of the oral airflow waveform may recover properties of the voice source excitation closer than that estimated from the acoustic waveform.

2. METHOD

This section describes the data collection from speakers with normal voices and two analysis methods: 1) GCI and GOI detection using YAGA [4] and 2) covariance method of linear prediction on acoustics and oral airflow to provide illustrations of closed phase inverse filtering performance.

2.1. Database

Ten adult speakers (five male, five female) with no history of voice disorders were recruited to participate in the study. The speakers were instructed to sustain five cardinal vowels (/a/, /e/, /i/, /o/, /u/) for 2–5 s at a comfortable pitch and loudness. Fig. 1 shows a snapshot of the three simultaneous signals recorded. The acoustic speech signal, oral airflow, and electroglottography (EGG) were sampled at 20 kHz. A circumferentially-vented pneumotachograph mask system (model MA-1L; Glottal Enterprises, Syracuse, NY) yielded the oral airflow signal.

2.2. Glottal closure and opening instant detection

GCIs and GOIs were extracted using YAGA [4] on the acoustic speech signal and the *derivative* of the oral airflow waveform. The derivative operation approximates the radiation characteristic to make the waveform suitable for epoch extraction. GCIs and GOIs were also detected from the derivative of the EGG signal as positive and negative peaks, respectively.

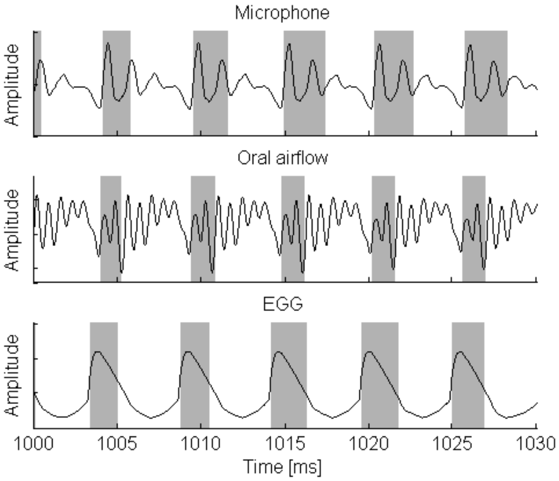

Fig. 1. Closed phases (shaded) identified in a segment of the sustained production of the /a/ vowel. The extracted GCIs and GOIs mark the start and the end, respectively, of the closed phase.

Figure 1 shows an exemplary 30 ms segment from the sustained vowel of one of the female speakers with shading representing the duration of the closed phase between the GCI and GOI within each cycle. The acoustic propagation time from the glottis where the EGG signal is recorded to the microphone is approximately 0.9 ms, the time delay between the GCIs detected from the EGG signal and the corresponding GCIs in the oral airflow and microphone signals. Fig. 1 also shows the difference in the closed phase estimation between the microphone, oral airflow and EGG signals, indicated by the width of the shaded area.

Table 1. Performance assessment of glottal closure and opening instant detection using acoustic microphone (Mic.) and oral airflow (Flow) recordings of five sustained vowels. IDR = identification rate, MR = miss rate, FAR = false alarm rate, ACC = accuracy.

		IDR [%]	MR [%]	FAR [%]	GCI ACC [ms]	GOI ACC [ms]
Mic.	Male	98.8	0.60	0.60	0.10	1.06
	Female	99.7	0.03	0.00	0.10	0.65
	All	99.4	0.40	0.20	0.10	0.79
Flow	Male	93.4	0.50	5.90	0.27	2.16
	Female	98.1	0.93	0.93	0.27	0.69
	All	96.8	0.81	2.35	0.28	1.42

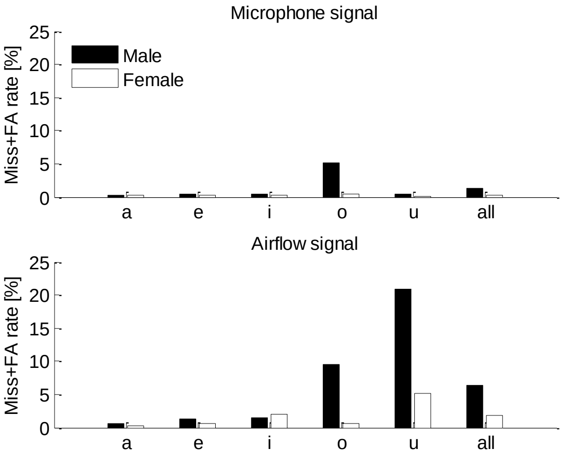

Fig. 2. Sum of miss and false alarm rates for GCI estimates from the microphone and airflow signals categorized by vowel.

An evaluation of the GCI and GOI estimates of the acoustic microphone and oral airflow waveforms was carried out using the EGG as the reference signal. Performance measures across all glottal cycles consisted of the identification rate, miss rate, and false alarm rate, and identification accuracy defined in [17]. A GCI is considered to be “identified” when the algorithm obtains a single GCI within a pitch period of a GCI in the reference signal. A miss occurs when there is no identified GCI within the reference pitch period, and a false alarm occurs when more than one GCI is identified within the reference pitch period. The identification accuracy is the standard deviation of the time differences between identified GCIs and their associated reference GCIs.

2.3. Closed phase inverse filtering

Exemplary results from applying the covariance method of linear prediction to obtain inverse filtered waveforms are presented. Autoregressive (AR) parameters are estimated over 25 ms analysis windows every 10 ms. The closed phase of each glottal cycle is defined as being bounded by a GCI and successive GOI extracted from the signal being analyzed. The first 0.2 ms of the closed phase are skipped to reduce the chance that the source excitation near the closure instant is included in the analysis. The AR parameters are then derived from all the closed phase samples that lie within each analysis window [18]. The signal is then inverse filtered by using the AR parameters as FIR filter coefficients to obtain the voice source signal.

3. RESULTS

Results are presented for GCI/GOI detection with EGG as reference (182,563 cycles) followed by exemplary inverse filter analysis of acoustic and oral airflow waveforms.

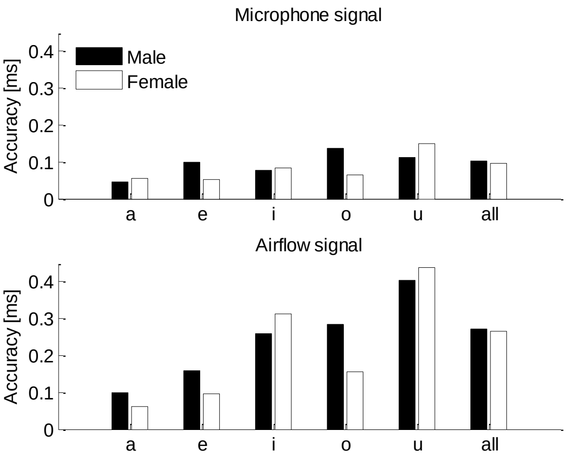

Fig. 3. Identification accuracy for GCI estimates using the speech signal and the oral airflow signal.

3.1. GCI and GOI detection

Table 1 reports the performance of GCI and GOI detection categorized by gender and across all speakers for the microphone and oral airflow recordings of the five vowels in the database. Since the identification of the GOIs is directly linked to the GCIs by allowing only one GOI per GCI, the same identification rate, miss rate, and false alarm rate result for both GCI and GOI performance. The difference in the timing accuracy is also reported.

For the microphone signal, the identification rate of 99.4 % is comparable to that found in a similar analysis on the APLAWD database, which also consists of 5 male and 5 female speakers [4]. The GCI timing accuracy of 0.10 ms on the acoustic signal is better than the 0.39 ms metric reported in [4]; but the GOI timing accuracy of 0.79 ms in the current study is lower than the 0.63 ms metric in [4]. For the oral airflow, the identification rate of 96.8 % is somewhat less than that for the microphone signal with the GCI timing accuracy of 0.28 ms remaining quite low. The timing accuracy of 2.16 ms of the GOI is, however, significantly higher for the male voices than for the female voices.

Figure 2 breaks down the GCI results to show gender-dependent miss and false alarm rates for all the vowels in the database. For the microphone-derived GCIs, the miss and false alarm rates remain particularly low for all vowels at a 1.25 % for males and 0.28 % for females (0.55 % combined); although performance is slightly degraded for /o/ by the males. The GCIs derived from the airflow, however, have much higher miss and false alarm rates, mostly due to the vowels /o/ and /u/ uttered by the male voices. Further investigation into vowel-dependent performance is warranted, since the results are not degraded in the acoustic speech signal.

Figure 3 breaks down the high identification accuracy (all below 0.5 s) from the microphone and the airflow signal. The best performance is found for the vowels /a/ and /e/, and the GCIs derived from the microphone signal gives

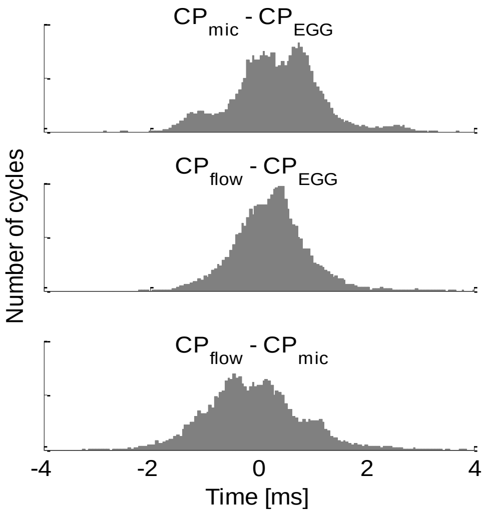

Fig. 4. Histograms of the pairwise differences between closed phase duration estimates (CP) of microphone and EGG (top), airflow and EGG (middle), and airflow and microphone (bottom).

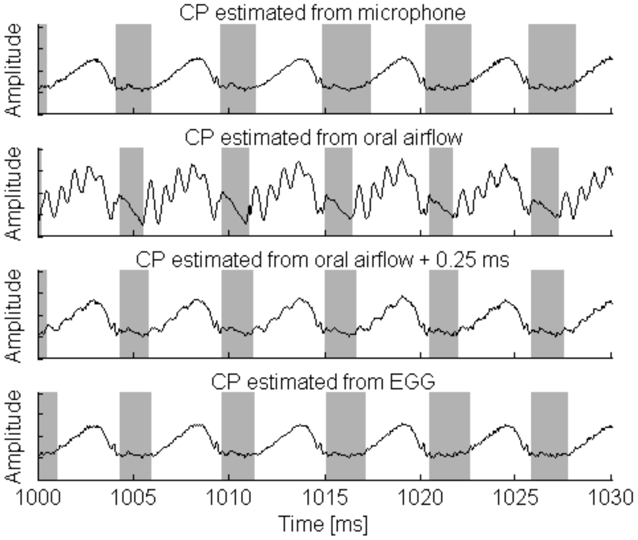

Fig. 5. Inverse filtered oral airflow waveforms for the adult female speaker uttering the vowel /a/. The closed phase (CP; shaded regions) timing is estimated in four different manners.

somewhat better accuracy than the GCIs derived from the oral airflow signal.

Figure 4 shows pairwise histograms of differences between the estimates of the closed phase duration from the microphone, EGG, and oral airflow signals. With the EGG as a reference for the closed phase duration, the top and middle panels show the deviation of the closed phase duration for the microphone and oral airflow signals respectively. From these distributions, both signals tend to overestimate the closed phase duration. The microphone-derived closed phase overestimates the EGG-based closed phase by an average of 0.33 ms, whereas the airflow-based closed phase overestimates by an average of 0.24 ms. The standard deviation of each distribution is quite large, however, at 0.80 ms for the microphone signal and 0.68 ms for the oral airflow. The airflow-based closed phase appears to give more consistent results with respect to the EGG-derived reference.

3.2. Closed phase inverse filtering

Figure 5 shows the inverse-filtered oral airflow waveform estimates for the same vowel in Fig. 1. Panel 1 and 4 show the oral flow inverse filtered using the closed phase (shaded) derived from the microphone and the EGG signal, respectively. Panel 2 shows the inverse filtered oral airflow using the closed phase derived from the oral airflow itself. In Panel 3, the closed phase was extended by adding 0.25 ms to the derived GOIs of Panel 2 to obtain a closed phase closer to that derived from the microphone and EGG. The reason for this addition was to show how being too conservative (too small a closed phase) can affect the covariance analysis, as shown in Panel 2 of Fig. 5. The inverse filtered oral airflow using its own GOIs does not perform as well as if the GOIs were taken from one of the other two channels (microphone or EGG). Panel 3 demonstrates that it is possible to rely on the GCIs extracted from the oral flow for the inverse filtering operation. The duration of the analysis period has to be long enough for first formant frequency estimation in covariance analysis.

These voice source signals offer the chance to derive important voice source features that are not easily derived from the acoustic waveform. For example the minimum flow can be estimated form these waveforms along with the maximum flow declination rate. In case of the microphone signal, the lip-radiation characteristic (typically considered to be a derivative operation) is still present in the inverse filtered signal and must be reversed. The inverse-filtered oral airflow, on the other hand, does not need to be integrated as the lip-radiation does not affect it, and the glottal airflow volume velocity is obtained directly.

4. CONCLUSION

Reliable identification of GCIs and GOIs from the acoustic speech signal has made closed-phase covariance analysis practicable without the use of an EGG signal or manual methods to obtain the closed phase [4, 5]. This paper provides initial results supporting the application of GCI detection on the oral airflow waveform that could potentially be applied to disordered voice analysis. GOI detection on the oral airflow waveform, however, compares less favorably than that computed on the acoustic microphone signal. Since YAGA was tuned to acoustic analysis, future work warrants a sensitivity analysis of parameters (number of poles, pre-filtering settings, etc.) to analyze oral airflow signals that do not include the lip-radiation characteristic.

5. REFERENCES

[1] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, “Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering,” J. Acoust. Soc. Am., vol. 125, no. 5, pp. 3289–3305, May 2009.

[2] P. Alku, “Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications,” Sādhāna: Indian Academy of Sciences Proceedings in Engineering, vol. 36, no. 5, pp. 623–650, 2011.

[3] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, "Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration," IEEE Trans. Audio Speech Lang. Processing, vol. 21, no. 9, pp. 1929–1939, 2013.

[4] 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. Acoust., Speech, Signal Process., vol. 20, no. 1, pp. 82–91, January 2012.

[5] T. Drugman, M.R.P. Thomas, J. Gudnason, P.A. Naylor, and T. Dutoit, “Detection of glottal closure instants from speech signals: A quantitative review,” IEEE Trans. Audio Speech Lang. Processing, vol. 20, no. 3, pp. 994–1006, March 2012.

[6] J. Sundberg, E. Fahlstedt, and A. Morell, “Effects on the glottal voice source of vocal loudness variation in untrained female and male voices,” J. Acoust. Soc. Am., vol. 117, no. 2, pp. 879–885, 2005.

[7] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” IEEE Trans. Biomed. Eng., vol. 55, no. 1, pp. 96-107, January 2008.

[8] A. Tsanas, M. A. Little, P. E. McSharry, L. O. Ramig, “Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests,’ IEEE Trans. Biomed. Eng., vol. 57, no. 4, pp. 884–893, April 2010.

[9] M. Rothenberg, “A new inverse filtering technique for deriving the glottal airflow waveform during voicing,” J. Acoust. Soc. Amer., vol. 53, pp. 1632–1645, 1973.

[10] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, “Objective assessment of vocal hyperfunction: An experimental framework and initial results,” J. Speech Hear. Res., vol. 32, no. 2, pp. 373–392, 1989.

[11] D.E. Veeneman and S.L. BeMent, “Automatic glottal inverse filtering from speech and electroglottographic signals,” IEEE Trans. Acoust., Speech, Signal Process., vol. 33, pp. 369–377, April 1985.

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

[13] H. W. Strube, “Determination of the instant of glottal closure from the speech wave,” J. Acoust. Soc. Am., vol. 56, no. 5, pp. 1625–1629, 1974.

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

[15] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal airflow derivative waveform with application to speaker identification,” IEEE Trans. Speech Audio Process., vol. 7, no. 5, pp. 569–576, Sept. 1999.

[16] D. Rudoy, T. F. Quatieri, and P. J. Wolfe, “Time-varying autoregressions in speech: Detection theory and applications,” IEEE Trans. Audio Speech Lang. Processing, vol. 19, no. 4, pp. 977–989, May 2011.

[17] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” IEEE Trans. Speech Audio Process., vol. 15, no. 1, pp. 34–43, January 2007.

[18] L. R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1978.