Interspeech 2018
2-6 September 2018, Hyderabad

Estimation of the asymmetry parameter of the glottal flow waveform using the electroglottographic signal

João P. Cabral

The ADAPT Research Centre, Trinity College Dublin, Ireland
cabralj@scss.tcd.ie

Abstract

Glottal activity information can be very important in several speech processing applications, such as in speech therapy, voice disorder diagnosis, voice transformation and text-to-speech synthesis. However, the use of algorithms for estimating glottal parameters from the speech signal is very limited in those applications because of problems with robustness and accuracy. For this reason, current research studies of the glottal source are usually constrained to isolated speech sounds or short segments of speech recorded in controlled conditions and methods requiring manual intervention. An alternative way to obtain more accurate and reliable glottal parameter estimates is to use other recording equipment besides the audio microphone. Electroglottography is the most popular non-invasive measurement of vocal fold motion. It has been widely used to estimate the glottal opening and closing instants, but it does not provide direct information about the other important glottal parameters. This paper proposes an automatic method for estimation of the glottal parameters from the electroglottographic signal that permits to measure an additional parameter related to the asymmetry of the glottal flow pulse. This is a very important characteristic correlated with voice quality and widely studied in voice source analysis, commonly represented by the speed quotient parameter.
Index Terms: glottal source analysis, speed quotient, electroglottography, EGG signal, voice quality, voice source

1. Introduction

A simple way of representing the speech production mechanism is by passing a source signal through a filter representing the vocal tract system and modelling the lip radiation by the first derivative. For voiced speech, the source signal carries very important information related to voice characteristics such as hoarseness, breathiness, creakiness, etc. [1, 2, 3, 4]. However, there are several problems in estimating the glottal source parameters from the speech signal [5, 6]: voiced/unvoiced classification errors, F0 estimation errors, difficulty to obtain an accurate estimation of the glottal source component from speech, and errors in parameterisation of the glottal source.

It has been shown that by using an Electroglottograph to measure the impedance variations of the larynx, it is possible to alleviate the voicing and F0 errors, for example in [5, 7, 8, 9] . The Electroglottographic (EGG) signal also permits to estimate accurately two important time instants of the glottal cycle: the glottal opening and closing instants [5, 7, 10, 11, 12, 8, 13]. In addition, the two-channel speech analysis can be used to improve glottal source signal estimation by using a pitch-synchronous inverse filtering technique, in which the analysis windows are centered at the glottal closure instants (GCIs) [5]. The EGG signal has been widely used to calculate the open phase duration of the glottal pulse, but in general it is not used to obtain the other important time instants of the glottal flow. This limitation is related to the fact that these two signals have different physical meaning, one representing an electric impedance while the other an acoustic signal, so they can not be easily mapped one to another analytically. This work studies the correlation between the EGG signal and an important characteristic of the glottal flow: the asymmetry of the glottal pulse shape. In order to achieve this, first a method is proposed to estimate as accurately as possible an asymmetry parameter on both the EGG signal and glottal source estimate. The hypothesis is that the asymmetry of the two signals is correlated so that the mapping can be derived through regression analysis of the relevant features.

The asymmetry of the pulse can be measured by extrating additional information about the closing phase (part of the open phase when the amplitude flow decreases). The work in [14] analysed relative measures of the opening and closing phases between the glottal flow and EGG signals. Measures from the EGG showed similarities to the airflow data and results support the hypothesis that the EGG signal has the ability to capture information about the asymmetry characteristic. The measurement criterion were based on segmenting the waveforms at different AC amplitude levels, by using a threshold between 20% and 80% of the waveform amplitude. However, such measures do not provide accurate details about the instants on the waveforms, which are extremely useful. For example, speech synthesis and voice transformation methods that use a glottal source model require accurate estimation of the parameters. Previous work can be found that measured the closing phase duration from the EGG signal. In [15], this measurement was used to compare an asymmetry parameter, the Speed Quotient (SQ), during modal and vocal fry phonation for different gender and vowel contexts. However, no papers were found that report results of the correlation between SQ measured on speech and EGG signals.

This work studies the correlation of a glottal flow asymmetry parameter between the EGG signal and the glottal flow, which is described in the next section. A method is proposed in Section 3 to accurately and robustly measure that parameter on both signals. The experiments in Section 4 show that there is significant correlation of the pulse asymmetry of the two signals and that the flow asymmetry can be predicted from the EGG signal only. Finally, the concluding remarks are given in Section 5.

2. Asymmetry Parameter of the Glottal Source

The glottal flow cycle can be divided into the closed phase (when the vocal folds are closed) and the open phase (when they are opened). The periodic variation in the air pressure gives the source energy for the voiced sounds. Figure 1 a) shows a segment of the glottal flow waveform $u_g(t)$ with cycle duration

10.21437/Interspeech.2018-2371

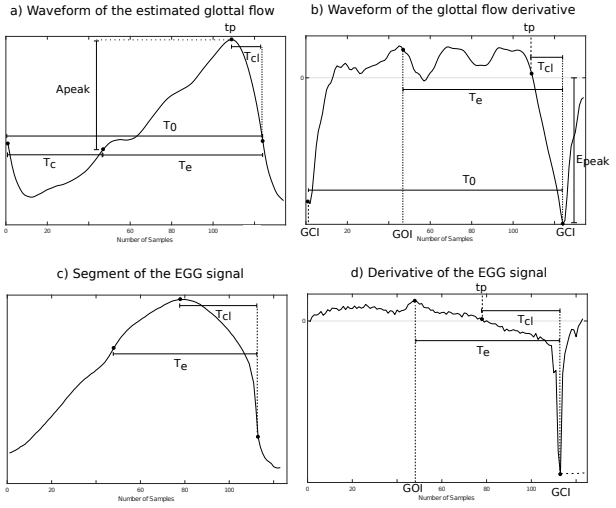

Figure 1: *Example of glottal parameter durations estimated on the glottal source and EGG signals.*

T_0, obtained by using a state-of-the-art glottal source estimation method called Iterative Adaptive Inverse Filtering (IAIF) technique [16]. Its derivative is also represented in Figure 1 b). The closed phase, with duration T_c, corresponds to the part of the glottal cycle when the energy is lower. Meanwhile, the open phase consists of the the peak flow pulse, with duration T_e, which starts at the glottal opening instant (GOI) and ends at the glottal closure instant (GCI). The GCI is also called epoch and corresponds to the instant of maximum excitation of the $u_g(t)$ derivative. The part of the pulse when the vocal folds are closing has duration T_{cl} and starts at the maximum of the glottal flow instant t_p.

Two of the most important dimensionless parameters of the glottal flow are the speed quotient (SQ) and open quotient (OQ). The first is related to the skewness of the flow pulse and is represented by $(T_e - T_{cl})/T_{cl}$. The OQ measures the relative duration of the open phase to the pitch period and can be represented by $OQ = T_e/T_0$. A more detailed explanation of their physical meaning can be found in [1, 4].

In this work, the correlation analysis is done by using another parameter related to the flow asymmetry called closing quotient, $CQ = T_{cl}/T_0$, which only depends on the glottal closure and t_p. Note that CQ is also often used to refer to the closed phase parameter that is equal to the inverse of OQ. The purpose of using CQ compared with SQ here is to avoid the effect of glottal opening estimation errors on the correlation. In particular, there are robust methods to estimate the glottal epochs on the speech and EGG signals, while the glottal opening estimation is generally more prone to errors [9, 17].

3. Analysis Method

3.1. Estimation of glottal opening and closing instants

Electroglottography is a very popular non-invasive measurement of vocal fold vibration. The EGG signal gives accurate information about the glottal closure (when the impedance is minimal) and glottal opening instants (when it is maximal). Figures 1 c) and d) show an example of the EGG signal and its derivative. The durations T_0 and T_{cl} of the glottal pulse can be robustly estimated from this signal, by detecting the opening and closing instants. Basically, the glottal closure can be obtained by detecting the strongest negative peak in the derivative of the EGG (DEGG) and the highest positive peaks on the DEGG for the opening instant. These two instants are also represented in Figures 1 c) and d). In this example, the signal has negative polarity according with VFCA (vocal-fold contact area) convention, that is, the polarity is positive when the EGG increases with glottal contact.

In this work, the GCI and GOI are estimated on the EGG signal by using the implementation of the peak detection algorithm called *peakdet* [18] available in the *covarep* toolbox [19]. *Peakdet* permits to choose different options for peak selection when multiple peak candidates are found in each analysis segment (corresponds to a pitch cycle). This effect makes the estimation problem more difficult, especially for the detection of the glottal opening instant [13]. The algorithm options are chosen by preliminary visual inspection of results on some EGG segments of the dataset. Before estimation with *peakdet*, the EGG signal is high-pass filtered in order to remove low frequency fluctuations with a linear-phase filter that has cut-off frequency of 50 Hz. Afterwards the GCIs that do not satisfy the minimum limit of fundamental frequency (F_0) for a given speaker are removed together with the opening instants of the same analysis frame (the *peakdet* algorithm already verifies the maximum F0).

For estimation of the glottal parameters from the speech signal $s(t)$, the glottal source estimation is performed using the IAIF method [16]. The GCIs are then estimated from the speech signal by using the RAPT algorithm algorithm [20] and [21] implemented in ESPS tools. In order to take advantage of the robust estimation of the GCIs and GOIs with the EGG signal, the epochs estimated from speech and EGG are aligned using an algorithm developed in previous work [22]. After this process, the epochs that are not aligned are removed as well as those that do not satisfy constraints on the pitch period (within the range of 50 Hz to 500 Hz). Note that there is a small delay between the EGG and speech signal which depends on the distance of the speaker's mouth to the microphone. Since this delay may not be constant, the GCIs from the EGG are not used as estimates of the GCIs of speech. The one-to-one mapping of pitch cycles obtained with the alignment is also required for the comparison of the CQ between the two signals. Meanwhile, the GOIs for speech (GOI^{sp}) are obtained by using the duration T_e calculated from the EGG signal, that is $GOI^{sp} = GCI^{sp} - T_e^{egg}$).

3.2. Estimation of maximal flow instant

A common way of estimating the instant of maximum flow t_p is by detecting the maximum of the glottal flow in the region delimited by the opening and closing instants, e.g. [17, 11, 7]. However, this peak detection method may not give a reliable result if the glottal flow pulse has multiple major peaks.

A method is developed in this work with the goal of improving the robustness of CQ analysis. It consists of first detecting the instant of maximum amplitude in the open phase and then performing a step to correct for eventual errors. This is done by detecting a CQ discontinuity with a threshold of the ΔCQ measured for two consecutive voiced frames. Then, the initial estimate t_p^a is compared with other candidate values to choose the one that gives the lowest delta value.

In order to obtain the candidate values, a new parameter is derived here based on the normalised amplitude quotient (NAQ) proposed in [23, 24]. NAQ uses the approximation that the glottal flow is represented by a triangular pulse that starts from zero at the glottal opening, reaches its maximum A_{max} at t_p and returns to the zero level at the glottal closure. Due to the rect-

angular shape of the derivative of the triangular pulse during the glottal closing phase, T_{cl} is calculated by the ratio between A_{peak} and the amplitude of the negative amplitude E_{peak} of the rectangular pulse: $NAQ = A_{peak}/(T_0 E_{peak})$. However, NAQ only gives a CQ-related measure because the triangular shape is considerably different from the glottal flow computed from real speech, which can be seen in the example of Figure 1. Similarly to the case of the triangular pulse shape, the value A_{peak} equals the area of the negative part of the glottal source derivative during T_{cl}. This area can be better approximated by the area of a triangle that is half the rectangle used to calculate NAQ. Then, in this work CQ is approximated by:

$$CQ = \frac{2A_{peak}}{T_0 E_{peak}} \quad (1)$$

From this equation, the new estimate t_p^b is calculated by the amplitude ratio and by $t_p^b = GCI - (T_0 CQ)$. Finally, the algorithm selects the best of the following three candidate values: t_p^a, t_p^b and the average obtained from the previous two.

4. Experiments

4.1. Dataset

The data consisted of speech recordings and EGG measurements from the US English BDL (male) voice of the CMU ARCTIC Dataset [25]. The first ten sentences were used in the experiments. Although this is a small subset of the database, it provided a large amount of data points for the correlation analysis (1081 in total corresponding to the number of pitch cycles analysed). The study is limited to a single speaker, given time constraints to perform manual verification of the parameter analysis on more data. However, the analysis is currently being extended to cover more variability factors of the speech signal, including speaker's gender.

4.2. Glottal parameter estimation

In the EGG analysis using the *peakdet* algorithm for GCI and GOI estimation, the different peak selection options (highest peak, first peak, last peak and barycenter methods) were compared by the author by manually verification of the results on some utterances. The highest peak method seemed to produce the most consistent and accurate results and was chosen for analysis. *Peakdet* also permits to set a threshold to detect GCI peaks either manually or automatically. This threshold was adjusted manually for each sentence, in which the criteria was to discard GCIs detected in voicing transitions (voiced-unvoiced and unvoiced-voiced) when the peaks have low amplitude. The idea was to favour a lower missing rate than false alarm rate, because the second can cause high parameter discontinuities that affect the correlation analysis. In general, the number of GCIs that were not detected in the transitions was up to 3. A smoothing step of *peakdet* performed on the DEGG to detect GCIs was enabled but it was disabled for GOI detection. This decision was based on preliminary testing by the author. Finally, both the estimated GCIs and GOIs were manually verified and corrected by visualising the labels on *wavesurfer*. Few corrections were necessary and they were mainly for removing GCIs in voicing transitions and adjusting the GOI when there was a high discontinuity of CQ due to wrong choice of the amplitude peak.

In the case of speech analysis, the IAIF method was implemented similarly to [26]. First, the speech signal sampled at 16 kHz was high-pass filtered at 50 Hz to remove the DC component. The signal was also downsampled to 4 kHz in order to better model the low frequency part of the glottal source. This is based on the known correlation of the OQ and SQ parameters with the low-frequency characteristics of the glottal source spectrum [4, 3]. Then, LPC inverse filtering was performed pitch-synchronously and iteratively to estimate the glottal source (consecutive LPC orders of 1 and 4) and the vocal tract (using the LPC order of 10 twice). The GCIs and GOIs were estimated from this signal as described in Section 3.1.

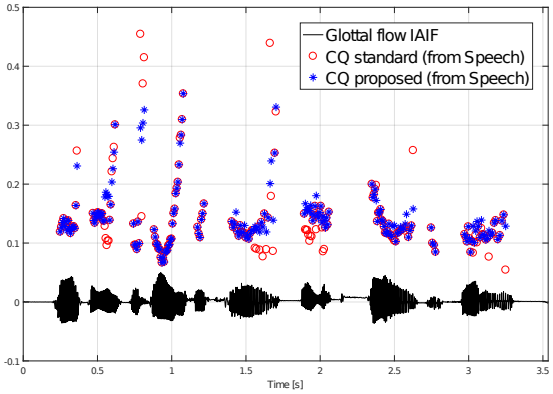

Figure 2: *CQ parameter trajectories estimated for a sentence glottal source signal obtained with the IAIF method.*

The instant of maximum amplitude t_p was obtained for the glottal flow and EGG using the method proposed in Section 3.1. The results were manually verified by using the *wavesurfer* software to visualise the glottal source waveform and the parameter labels. Figure 2 shows an example of the effect of using the amplitude quotient to reduce discontinuities. In general, t_p^b estimated using equation (1) gives a good approximation to the instant of maximal flow t_p^a, but it is assumed to be less accurate because it relies on the approximation of the triangle area. However, t_p^b helps to reduce the effect of outliers to obtain a smoother parameter contour.

Figure 2 shows that CQ can vary significantly between frames, which may be produced by natural action of the vocal folds or due to errors in parameter estimation. On the other hand, it was verified that the T_0 parameter trajectories were smooth, which reflected the good performance of GCI detection on both speech and EGG. Since the author observed that the automatic detection of the maximum amplitude peak is very robust for both EGG and glottal source signal analysis, possible errors could be related to the difference of these signals to the true glottal source signal. It was also observed that the CQ trajectories obtained with EGG were in general smoother than those obtained from speech. But this can not be used as an indicator that the EGG is more accurate because there is no direct relation between the EGG pulse shape and the glottal shape, e.g. as shown in [27].

4.3. Data analysis

The CQ was computed using the following methods:

- A_{speech}: Proposed method for CQ estimation from speech, based on the posteriori refinement of t_p with the amplitude ratio parameter.
- B_{speech}: Standard method for CQ estimation from speech (without using the amplitude ratio).
- C_{egg}: Method for estimation of CQ from the EGG.

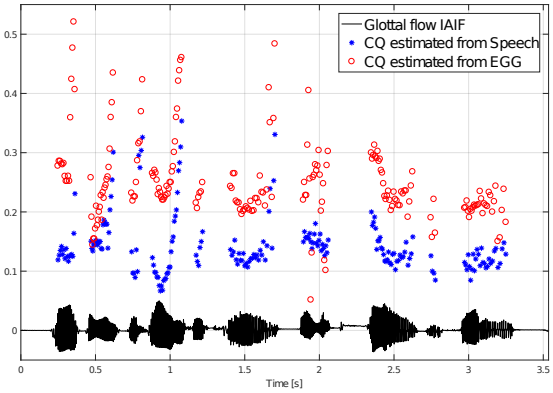

Figure 3: *Comparison of the CQ parameter trajectories estimated for a sentence from the speech signal and EGG.*

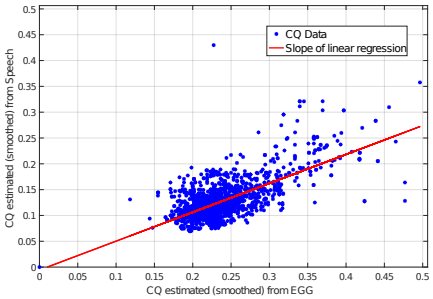

Figure 4: *Linear regression relation between CQ estimated from speech by the method A_{speech} and from the EGG signal.*

Then, the CQ values obtained for each method were smoothed using the median function with a window length of 5 samples, in order to reduce the effect of outliers on the correlation analysis.

The results of a Pearson's linear correlation test showed that there is a significantly high correlation between CQ estimated from speech using both speech analysis methods (A_{speech} and B_{speech}) and CQ estimated from the EGG signal. A_{speech} obtained a correlation value of 0.6827 ($p < 0.001$), while B_{speech} obtained the slightly higher coefficient of 0.7245 ($p < 0.001$). Figure 3 shows an example of the CQ parameter contours for a sentence obtained with the methods A_{speech} and C_{egg}.

Given the strong linear correlation obtained in the comparison of the A_{speech} with C_{egg} methods, a simple linear regression was performed to determine how well a linear model can be used to predict the values of A_{speech} using C_{egg}. Figure 4 shows that by plotting the data of the two methods for the BDL voice subset there is a clear linear correlation, represented by the line with slope 0.56 that crosses the *y* axis at -0.006.

The linear model was used to predict the values of CQ estimated on the glottal flow from the CQ estimated from the EGG signal. Figure 5 shows an example of the CQ trajectory estimated from the flow waveform and the one obtained from the EGG signal using this model. This example shows that CQ values predicted from EGG can fit very well to the values of A_{speech}, but in some regions the fit is not as good. One possible explanation is that it may be necessary a more sophisticated model, such as a higher-order polynomial model or nonlinear model, to obtain a better fitting. Other factor that may contribute to improve the fitting is to consider other features related to the pulse shape which are also correlated with CQ. For example, the author also tested the correlation between the OQ estimated from the EGG signal and CQ estimated using the A_{speech} and

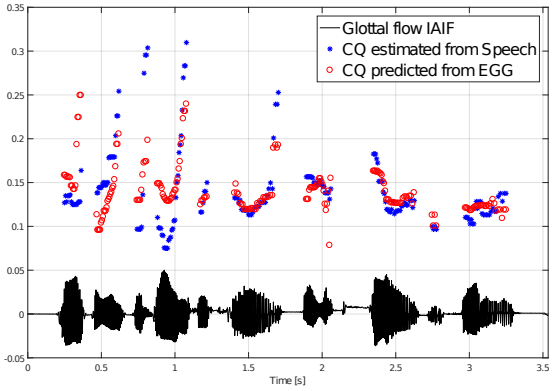

Figure 5: *Contours of smoothed CQ estimated from speech using the proposed method A_{speech} and from the EGG signal using the linear model.*

C_{egg} methods. Results showed that there is a significant correlation ($p < 0.001$) between the CQ feature of C_{egg} (correlation coefficient equal to 0.674) and the observation CQ obtained with A_{speech} (coefficient equal to 0.669). The mean squared error (MSE) between the CQ obtained with the method A_{speech} and the CQ estimated from the EGG signal using the linear model was 0.000735. This error was much lower than the error calculated between A_{speech} and C_{egg}, which was equal to 0.0134. Thus, the linear transformation of the CQ values obtained with the C_{egg} method produces a much better approximation to the CQ estimated from the glottal flow signal.

The results of this preliminary experiment indicate that the EGG signal can be used to estimate the asymmetry parameter of the glottal flow, either using the EGG only, e.g. if the speech signal is noisy or other reason that makes the analysis difficult. Other application of this model is in a method that can improve the estimation of the glottal flow asymmetry using the 2-channel input. These are only initial results from an ongoing experiment and currently more data is being analysed. Also, the performance of CQ prediction by the EGG needs to be better evaluated using an held-out test set.

5. Conclusions

This paper studied the analysis of a glottal parameter related to the symmetry of the glottal flow signal, the CQ parameter. A method was proposed to robustly estimate the CQ parameter on the glottal flow signal, by combining the output of a peak detection algorithm with an amplitude ratio estimate. By performing measurements of this parameter on the speech and the EGG signal it was found that there is a high correlation of the parameter estimates between the two signals. Moreover, the preliminary experiment also showed that the CQ estimated from the EGG signal can be transformed using a linear model to obtain CQ values more similar to those estimated from the glottal flow waveform. These results motivate the ongoing work of the author on extending this experiment to more data and more complex regression analysis models.

6. Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT centre (www.adaptcentre.ie) at Trinity College Dublin and the Irish Research Council funded Cog-SiS project (Grant R17339).

7. References

[1] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, 1995.

[2] E. Keller, "The analysis of voice quality in speech processing," *Lecture notes in computer science*, vol. 3445, pp. 54–73, 2005.

[3] C. Gobl, "A preliminary study of acoustic voice quality correlates," Royal Institute of Technology, KTH, Stockholm, STL-QPSR, 1989.

[4] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," Royal Institute of Technology, KTH, Stockholm, STL-QPSR, 1995.

[5] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Trans Signal Processing*, vol. 34, pp. 730–743, Feb. 1986.

[6] T. Drugman, M. R. P. Thomas, J. Gunason, P. A. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 994–1006, 2012.

[7] D. Thotappa and S. R. M. Prasanna, "Reference and automatic marking of glottal opening instants using egg signal," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, 2014, pp. 1–5.

[8] A. Bouzid and N. Ellouze, "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," *Speech Communication*, vol. 51, no. 9, pp. 782 – 792, 2009, special issue on non-linear and conventional speech processing.

[9] N. Sturmel, C. d'Alessandro, and B. Doval, "A spectral method for estimation of the voice speed quotient and evaluation using electroglottography," 2006.

[10] K. Ramesh and S. R. Prasanna, "Glottal opening instants detection using zero frequency resonator," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 127–141, 2017.

[11] K. Ramesh, S. R. M. Prasanna, and D. Govind, "Detection of glottal opening instants using hilbert envelope," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 44–48.

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[13] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation." *The Journal of the Acoustical Society of America*, vol. 115 3, pp. 1321–32, 2004.

[14] C. M. Sapienza, E. T. Stathopoulos, and C. Dromey, "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms : Effects of measurement criteria and sound pressure level," *Journal of Voice*, vol. 12, no. 1, pp. 31 – 43, 1998.

[15] Y. Chen, M. P. Robb, and H. R. Gilbert, "Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 5, pp. 821–829, 2002.

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

[17] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal closing and opening instants using an improved dypsa framework," in *2009 17th European Signal Processing Conference*, 2009, pp. 2191–2195.

[18] M. Mazaudon and A. Michaud, "Tonal contrasts and initial consonants: A case study of tamang, a 'missing link' in tonogenesis," *Phonetics*, vol. 65, no. 4, pp. 231–256, 2008.

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[20] D. Talkin and J. Rowley, "Pitch-synchronous analysis and synthesis for TTS systems," in *Proc. of ESCA Workshop on Speech Synthesis*, Autrans, France, September 1990.

[21] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, 1995, pp. 495–518.

[22] J. P. Cabral, J. Kane, C. Gobl, and J. Carson-Berndsen, "Evaluation of glottal epoch detection algorithms on different voice types," in *INTERSPEECH*. International Speech Communication Association (ISCA), 2013, pp. 1989–1992.

[23] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, pp. 701–710, 2002.

[24] J. Lohscheller, J. G. Švec, and M. Döllinger, "Vocal fold vibration amplitude, open quotient, speed quotient and their variability along glottal length: Kymographic data from normal subjects," *Logopedics Phoniatrics Vocology*, vol. 38, no. 4, pp. 182–192, 2013.

[25] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop (SSW5)*, Pittsburgh, USA, 2004.

[26] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis," in *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, August 2007, pp. 113–118.

[27] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 666–677, 2016.