*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Classification of Voice Modality using Electroglottogram Waveforms

*Michal Borsky*[1], *Daryush D. Mehta*[2], *Julius P. Gudjohnsen*[1], *Jon Gudnason*[1]

[1] Center for Analysis and Design of Intelligent Agents, Reykjavik University
[2] Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital, Boston, MA
michalb@ru.is, mehta.daryush@mgh.harvard.edu, juliusg15@ru.is, jg@ru.is

## Abstract

It has been proven that the improper function of the vocal folds can result in perceptually distorted speech that is typically identified with various speech pathologies or even some neurological diseases. As a consequence, researchers have focused on finding quantitative voice characteristics to objectively assess and automatically detect non-modal voice types. The bulk of the research has focused on classifying the speech modality by using the features extracted from the speech signal. This paper proposes a different approach that focuses on analyzing the signal characteristics of the electroglottogram (EGG) waveform. The core idea is that modal and different kinds of non-modal voice types produce EGG signals that have distinct spectral/cepstral characteristics. As a consequence, they can be distinguished from each other by using standard cepstral-based features and a simple multivariate Gaussian mixture model. The practical usability of this approach has been verified in the task of classifying among modal, breathy, rough, pressed and soft voice types. We have achieved 83% frame-level accuracy and 91% utterance-level accuracy by training a speaker-dependent system.

**Index Terms**: electroglottogram waveforms, non-modal voice, MFCC, GMM, classification

## 1. Introduction

The standard model of speech production describes the process as a simple convolution between vocal tract and voice source characteristics. In this model, the vocal tract is modeled as a series of passive resonators that provides phonetic context to speech communication. The voice source signal provides the driving signal that is modulated by the vocal tract. The process of creating the voice source signal is a complex process in which the stream of air exiting the lungs is passed through the vocal folds that open and close to modulate the air flow. Although the characteristics of the source signal are generally less complex than the output speech, it carries vital information relating to the produced speech quality.

There are several methods of analyzing the voice source separately from the vocal tract, including endoscopic laryngeal imaging, acoustic analysis, aerodynamic measurement, and electroglottographic assessment. Each approach yields slightly different results as different signals are utilized. For acoustic or aerodynamic assessment, the voice source signal is obtained through an application of inverse filtering that removes vocal tract-related information from the radiated acoustic or oral airflow signal [1]. For electroglottographic assessment, the objective is to analyze the patterns of vocal fold contact indirectly through a glottal conductance, or electroglottogram (EGG), waveform [2].

Subjective voice quality assessment has a long and successful history of usage in the clinical practice of voice disorder analysis. Historically, several standards have been proposed and worked with in order to grade the dysphonic speech. One popular auditory-perceptual grading protocol is termed GRBAS [3], which comprises five qualities - grade (G), breathiness (B), roughness (R), asthenicity (A), and strain (S). Another popular grading protocol is the CAPE-V [4] which comprises of auditory-perceptual dimensions of voice quality that include overall dysphonia (O), breathiness (B), roughness (R), and strain (S). These qualitative characteristics are typically rated subjectively by trained personnel who then relate their auditory perception of the voice to the associated laryngeal function.

The exact nature and characteristics of the non-modal voice types continues to be investigated. However, the general consensus is that the breathy voice type is characterized by an overall turbulent glottal airflow [5], the pressed voice type is associated with an increased subglottal pressure (as if voicing while carrying a heavy suitcase), and the rough voice type by temporal and spectral irregularities of the voicing source. Speech scientists, speech signal processing engineers, and clinical voice experts have been collaborating on developing methods for the automatic detection of non-modal phonation types. The bulk of research has focused on classification between pathological and normal speech has been extensively developed in recent years, see [6, 7, 8, 9, 10]. In contrast, the classification of voice mode represents a comparatively less developed research field. The authors in [11] employed a set of spectral measures (fundamental frequency, formant frequencies, spectral slope, H1, H2, H1-H2) and achieved 75% accuracy of classification between modal and creaky voice (a non-modal voice type associated with reduced airflow and temporal period irregularity). In another study [12], similar classification accuracy of 74% was reported for the task of detecting vocal fry. A task very similar to the one presented in this paper was explored in [13], where the authors used skin-surface microphones to indirectly estimate vocal function in order to classify laryngeal disorders, but ultimately concluded that acoustic information outperformed surface microphone information.

The current study proposes a different approach that focuses on analyzing vocal function indirectly by exploiting the frequency characteristics of EGG waveforms. The main objective of this paper is to present the results of this novel approach to automatic classify modal and different types of non-modal voice types. The paper is organized as follows. Section 2 provides a short overview of the nature of the EGG waveform. Sections 3 and 4 describe the experimental setup and the achieved results, respectively. The paper concludes with a discussion of future work in Section 5.



http://dx.doi.org/10.21437/Interspeech.2016-1194

## 2. Characteristics of the EGG signal

The electroglottograph is a device that was developed to monitor the opening and closing of the vocal folds, as well as vocal fold contact area, during phonation. The device operates by measuring the electrical conductivity between two electrodes that are placed on the surface of the neck at the laryngeal area. The output EGG waveform correlates with vocal fold contact area; thus, the EGG signal is at its maximum when the vocal folds are fully closed, and the EGG signal is at its minimum when the folds are fully opened [2]. The instants of glottal opening and closure are most prominent during modal phonation but can often be observed even during soft and breathy speech depending on the degree of vocal fold contact.

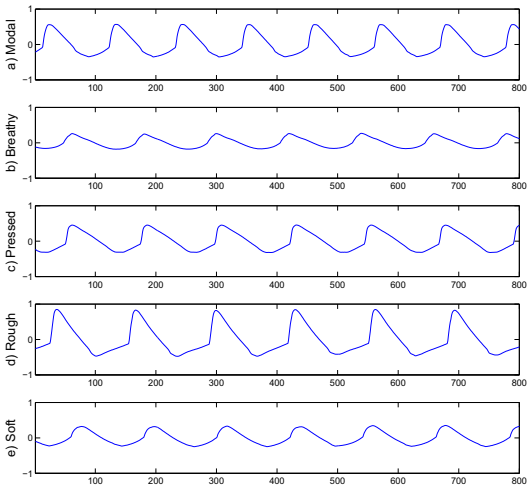


Figure 1: *Characteristic EGG waveforms of modal and 4 types of non-modal voice types.*

Throughout the years, researchers have demonstrated that the periodic vibrations of the vocal folds correlate with the characteristic shape of the EGG waveform [14, 15, 16]. These attributes are usually exploited to better understand vocal fold contact characteristics. Another popular EGG application is the detection of glottal closure instants (GCIs) and glottal opening instants (GOIs) using, e.g., the SIGMA algorithm in [17].

Figure 1 displays an example of five different voice types that were studied in this paper: modal, breathy, rough, soft, and pressed voice types. The principal idea is to use standard mel-frequency cepstral coefficient (MFCC) features extracted from the EGG signal and a Gaussian mixture model (GMM) to classify among modal and non-modal voice types. The hypothesis is that modal and different kinds of non-modal voice types produce EGG signals that have distinct spectral characteristics.

An example log-spectrum of the EGG waveform recorded from a vocally normal speaker producing modal phonation is illustrated in Figure 2. The spectrum is characterized primarily by peaks that correspond to the fundamental frequency and higher harmonic components. The spectrum decays rapidly while the majority of the information is carried by frequencies $\leq$ 4000 Hz. The experimental setup adopted in this study employs MFCC features extracted from the EGG signal. There were two main reasons for this. First, the MFCCs are a convenient and well established method to model the spectrum in a compact way. Second, the mel-frequency filter bank is most sensitive at lower frequencies, which is where most of the information is contained for the EGG waveform.

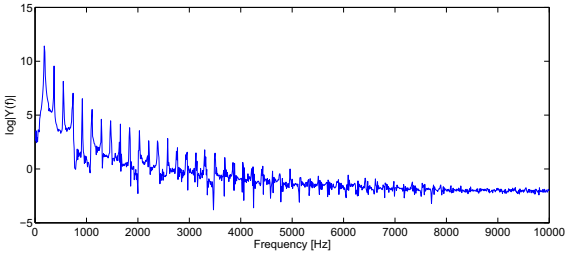


Figure 2: *EGG spectrum for modal speech.*

## 3. Method

### 3.1. Database

The experiments presented in this paper were performed on a database that contains recordings collected in an acoustically treated sound booth. The whole set consisted of 11 speakers (six males, five females) with no history of voice disorders and endoscopically verified normal vocal status. Each speaker produced several utterances of running speech and sustained vowel tasks. The participant were asked to produce the vowels in their typical (modal) voice and later in four different types of voice quality: breathy, pressed, soft, and rough. Elicited tokens were monitored by a speech-language pathologist; future work calls for the auditory-perceptual rating of the elicited tokens since it is challenging to produce a 'pure' non-modal voice type. Several other speech-related signals were recorded from each participant, which were later time-synchronized and amplitude-normalized. Some speakers read the utterance set only once, whereas others repeated tokens multiple times. All signals were sampled at $f_s = 20$ *[kHz]*. The experiments were performed with recordings of the sustained vowels '*a, e, i, o, u*'.

### 3.2. Experimental Setup

The process of constructing the classifier started with extracting the features. Parameters applied are as follow:

- Frame: Length = 2048 samples, Shift = 256 samples (87.5% overlap), Hamming window
- Mel-filter bank: 128 filters, $f_{min} = 50$ $[Hz]$, $f_{max} = 4000$ $[Hz]$
- Number of MFCCs: 14 (13 static MFCCs + $0^{th}$ coefficient)

This parametrization is very similar to what is generally used in automatic speech recognition systems, where the only notable differences were the frame-length and the number of filters in the Mel-frequency filter bank. The higher number of Mel-bank filters resulted in a higher spectral resolution, especially at lower frequencies. The frame-length used in our experiments was set to approximately 100 $[ms]$, which was justified due to the statistical quasistationarity for the sustained vowels in the database. Table 1 summarizes the total number of frames and the number of MFCC vectors for each voice type.

Table 1: *Number of frames for each voice type*

| Modal | Rough | Breathy | Pressed | Soft |
|---|---|---|---|---|
| 20 296 | 12 623 | 11 764 | 10 335 | 21 530 |

The constructed classifier was based on GMMs characterized by their full covariance matrices. The means of distributions for each class were initialized to randomly selected data points from that class. The model parameters were re-estimated in a supervised fashion using the expectation-maximization (EM) algorithm.

In order to draw statistically significant conclusions, we established two different classification setups. In the first case, one utterance was set aside as the test utterance while the rest of data was used to train the models. The process was then repeated for all signals in order to obtain a confusion matrix. This approach allowed us to evaluate the classification accuracy both at the utterance and the frame level. In the second case, all frames were pooled together regardless of their content and then randomly split into training-test sets with a 9:1 ratio. The process was repeated multiple (64) times to ensure results were robust to outlier performance. The purpose of this second setup was to avoid training content-dependent classifiers and to examine general effects of voice type on speech. However, this setup only allowed for evaluating frame-level classification accuracy.

## 4. Results and Discussion

This section summarizes the results from the series of classification experiments on the descriptive and discriminative qualities of the EGG signal. A detailed description of each classification setup is provided in the corresponding section.

### 4.1. Separability of voice types using the EGG signal

The accuracy of the classification task depends on extracting features that are capable of separating classes from each other in a given feature space. Figure 3 shows the spread of observations for all the non-modal voice types from one speaker in the MFCC[0]-MFCC[1] plane. Although the data points in this figure were obtained from a single speaker, there are still several interesting things to note. First, different voice types occupy different positions in the space, which certainly supports the assumption that distinct voice types can potentially be separated from each other using MFCCs. Second, breathy and soft voice types appear to overlap. This observation indicates that EGG spectra for these two voice types are similar (which was expected), and thus classification between breathy and soft phonation is challenging. Third, the pressed and rough voice types are located near each other while the modal voice is located in between. Finally, the outlier data points are in fact silence segments as no voice activity detection was applied to remove them. Rather, we set the number of mixtures to two and let the system model these ”garbage” frames with one mixture from each class. Although Figure 3 is a simplification of the analysis by only displaying the first two MFCCs, the exercise was instructive to begin to understand the separability of voice types using MFCCs of the EGG signal.

### 4.2. Two-class classification

In the first series of experiments, we constructed speaker-dependent classifiers that were trained and tested on data from a single speaker. The primary goal was to avoid introducing additional speaker variability and to measure the discriminative potential of MFCC features extracted from the EGG signal in the most optimal scenario. These experiments were performed using the second data splitting method.

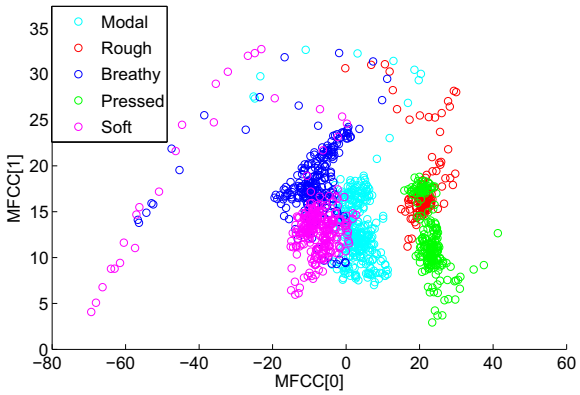


Figure 3: *Modal, rough, breathy, pressed and soft voice in MFCC[0]-MFCC[1] plane.*

Table 2 summarizes results from a two-class classification task between modal and one type of non-modal voice type. This setup excludes the potential of overlap among non-modal voice types and focuses solely on assessing the differences between modal and any manifestation of non-modal voice type. Even though the task is fairly simple, it is still able to provide an initial insight into the discriminatory qualities of EGG using objective methods to complement the observations of the scatter plot in Figure 3. The highest accuracy of 98.74% was achieved for the rough voice. These results would indicate that the rough voice type is easily distinguishable from modal speech. These results were followed closely by breathy, pressed, and soft voice types. The obtained results demonstrate that classification of modal and non-modal speech may be successfully accomplished using EGG waveforms.

Table 2: *Frame-level accuracy [%] of two-class classification between modal and a given non-modal voice type.*

| | Rough | Breathy | Pressed | Soft |
|---|---|---|---|---|
| Modal | 98.74 | 97.22 | 96.78 | 94.46 |

### 4.3. Frame-level splitting five-class classification

Whereas the purpose of the previous section was to do an initial evaluation on the separability of voice types using EGG, the goal of this section was to perform more realistic tests using five-class classifiers. The main advantage of this setup was the fact that it took potential overlap among different non-modal voice types into account. The data was once again split using the random frame distribution method. The frame-normalized accuracy for all speech types is summarized in the full confusion Table 3.

There are several interesting conclusions that can be drawn from Table 3. The modal voice type achieved the highest overall classification accuracy of 93.8% and was most often confused with soft and breathy voice, in that order. The second-best results were obtained for breathy voice (89.5%), followed by pressed (83.3%), rough (83.3%), and soft (79.5%) voice types. A closer analysis of the confusion table support the previ-

ously stated conclusions about data overlap to a certain degree. We observe that breathy voice is most often confused with soft speech (4.4%); however, the converse was not true. Soft voice frames were labeled as being pressed more often than breathy. Another interesting thing to note was the fact that a relatively wide spread of rough voice into other clusters caused problems for all other non-modal voice types; this result may be due to the intermittent and unstable production of a rough-sounding voice. These voice types were produced by untrained speakers, and it is highly probable that multiple voice types were exhibited in each token. Similar, the pressed voice type is difficult to elicit as a pure dimension and consequently contributes to its classification as either breathy or pressed.

Results support the conclusion from the previous experiment and prove that voice modality may be successfully identified solely from the EGG signal. The results also indicate that a $100[ms]$ segment is satisfactory to classify voice type with an average accuracy of 83%.

Table 3: *Frame-level accuracy [%] of five-class classification with data frames split randomly into training and test sets.*

| | | Recognized | | | | |
|---|---|---|---|---|---|---|
| | | Modal | Rough | Brea. | Press. | Soft |
| Actual | Modal | **93.8** | 1.2 | 1.7 | 0.8 | 2.5 |
| | Rough | 0.7 | **83.3** | 4.6 | 8.2 | 3.2 |
| | Brea. | 1.8 | 3.1 | **89.5** | 1.2 | 4.4 |
| | Press. | 1.7 | 7.9 | 3.0 | **83.3** | 4.1 |
| | Soft | 2.7 | 5.7 | 4.1 | 8.0 | **79.5** |

### 4.4. Utterance-level splitting five-class classification

Splitting data at the utterance level and assigning certain frames from the same utterance to both the training and test sets creates a problem as the classifiers are potentially able to learn on the test data. Due to this reason, the following five-class classification task was performed with data that was split at the utterance level. As a consequence, it allowed for the comparison of both frame-level and utterance-level classification accuracy.

Table 4 summarizes the frame-level five-class classification performance using the utterance level split. As such, these results are directly comparable to the ones already presented in Table 3. We can observe a general trend of declining accuracy for all voice types. The lowest performance drop of 1.34 percentage points (pp) was observed for soft speech. We saw a 4 pp drop for modal, rough, and pressed voice types and 13 pp for breathy. One interesting thing to note was the fact that breathy voices were misclassified as soft in approximately the same number of cases as soft was misclassified for breathy; 11.33% vs. 11.18%, respectively. Finally, rough and pressed voice types displayed qualitatively similar trends as they were often misclassified for each other. Our previous experiments did not display this kind of clear division between different voice types.

Table 5 summarizes the utterance-level accuracy that was obtained from the frame-level classification by selecting the most occurring class. Although we observe a significant increase in the overall accuracy across all classes, the general trends correspond to the trends observed in Table 3.

## 5. Conclusion and Future Work

This paper presents a novel approach of voice modality classification that is based on processing the EGG signal, an indirect measure of vocal fold contact available in laboratory settings. The EGG waveforms were parametrized using a standard MFCC scheme, and the extracted features were then classified using GMMs. The models were trained to be speaker dependent, and a series of tests were conducted to demonstrate the viability of this approach. The primary task was to classify among modal, breathy, rough, pressed, and soft voice types. The presented method achieved 83% frame-level accuracy and 91% utterance-level accuracy. A closer look at the confusion matrix reveals that modal voice achieved the highest accuracy regardless of the classification task and setup. This result indicates that the spectral composition of modal EGG is more distinct from other non-modal EGGs than the non-modal types are different from each other. The breathy voice type was observed to be similar to the soft voice type, and rough was often interchangeable with pressed voice. In fact, the reality is that the frames of a particular utterance may be characterized not only by multiple voice modes within the same token, but each frame may be described as exhibiting proportions of the different non-modal voice types. Auditory-perceptual ratings of an utterance along various dimensions (e.g., using the CAPE-V form) may aid in enhancing the ground truth labeling of voice type.

This work represents an initial study on the discriminatory qualities of EGG waveforms and their spectral characteristics for voice modality classification. Current results indicate that there is a great variation in EGG among speakers which makes construction the speaker-independent classifier a challenging problem at the moment. The authors believe that the described methods can be extended into the field of dysphonic speech classification as the studied qualities are often observed by patients with various voice pathologies. This clinical direction represents the potentially most important application of this work.

Table 4: *Frame-level accuracy [%] with taking out one utterance and training on rest.*

| | | Recognized | | | | |
|---|---|---|---|---|---|---|
| | | Modal | Rough | Brea. | Press. | Soft |
| Actual | Modal | **89.1** | 1.5 | 1.9 | 2.6 | 4.9 |
| | Rough | 0.7 | **78.5** | 7.1 | 11.4 | 2.3 |
| | Brea. | 3.6 | 6 | **73** | 6.1 | 11.3 |
| | Press. | 2.9 | 9.8 | 4.6 | **80.6** | 2.1 |
| | Soft | 6.0 | 2.3 | 11.2 | 2.3 | **78.2** |

Table 5: *Utterance-level accuracy [%] with taking out one utterance and training on rest.*

| | | Recognized | | | | |
|---|---|---|---|---|---|---|
| | | Modal | Rough | Brea. | Press. | Soft |
| Actual | Modal | **98.5** | 0 | 0 | 0 | 1.5 |
| | Rough | 0 | **85.5** | 2.7 | 11.8 | 0 |
| | Brea. | 1.7 | 1.7 | **91.4** | 1.7 | 3.5 |
| | Press. | 0 | 7.8 | 1.6 | **90.6** | 0 |
| | Soft | 2.9 | 0 | 5.7 | 0 | **91.4** |

## 6. Acknowledgments

This work is sponsored by The Icelandic Centre for Research (RANNIS) under the project *Model-based speech production analysis and voice quality assessment*, Grant No 152705-051. This work was also supported by the Voice Health Institute and the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders under Grant R33 DC011588. The papers contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

# 7. References


[1] P. Alku, "Eurospeech '91 glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992.

[2] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic Assessment of Normal Voice: A Tutorial," *Clinical Linguistics and Phonetics*, vol. 3, no. 3, pp. 263–296, 1989.

[3] H. Minoru and K. R. McCormick, "Clinical examination of voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 4, October 1986.

[4] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, , and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," *American Journal of Speech Language Pathology*, vol. 18, no. 2, pp. 124–132, May 2009.

[5] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[6] J. W. Lee, S. Kim, and H. G. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 5969–5973.

[7] S. N. Awan, N. Roy, M. E. Jett, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the cape-v," *Clinical Linguistics & Phonetics*, vol. 24, no. 9, pp. 742–758, 2010.

[8] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using mfcc and gmm," in *GCC Conference and Exhibition (GCC), 2013 7th IEEE*, Nov 2013, pp. 292–297.

[9] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[10] P. Henriquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. D. de Maria, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1186–1195, Aug 2009.

[11] T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson, "Detecting non-modal phonation in telephone speech," in *Proceedings of the Speech Prosody 2008 Conference*. Lbass, 2008. [Online]. Available: https://books.google.is/books?id=92urUXy8RJ8C

[12] C. T. Ishi, K. I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47–56, Jan 2008.

[13] A. Gelzinis, A. Verikas, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hllander, V. Uloza, and E. Padervinskis, "Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders," in *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*, Dec 2014, pp. 125–132.

[14] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992.

[15] D. Childers, D. Hicks, G. Moore, L. Eskenazi, and A. Lalwani, "Electroglottography and vocal fold physiology," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 245–254, 1990.

[16] C. Painter, "Electroglottogram waveform types," *Archives of oto-rhino-laryngology*, vol. 245, no. 2, pp. 116–121, 1988.

[17] M. Thomas and P. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1557–1566, Nov 2009.