Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 624575, 11 pages
doi:10.1155/2011/624575

Research Article

Spoken Emotion Recognition Using Glottal Symmetry

Alexander I. Iliev and Michael S. Scordilis

Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

Correspondence should be addressed to Alexander I. Iliev, ailiev@miami.edu

Received 1 August 2010; Revised 30 October 2010; Accepted 28 February 2011

Academic Editor: Julien Epps

Speech variability in real-world situations makes spoken emotion recognition a challenging task. While a variety of temporal and spectral speech features have been proposed, this paper investigates the effectiveness of using the glottal airflow signal in recognizing emotions. The speech used in this investigation is from a classical recording of the theatrical play "Waiting for Godot" by Samuel Beckett. Six emotions were investigated: happy, angry, sad, fear, surprise, and neutral. The proposed method was tested on the original recording and on simulated distortion conditions. In clean signal conditions the proposed method achieved average recognition rates of 76% for four emotions and 66.5% for all six emotions. Furthermore, it proved fairly robust under signal distortion and noisy conditions achieving recognition rates of 60% for four and 51.6% for six emotions for severely low-pass filtered speech, while with additive white Gaussian noise at SNR = 10 dB recognition rates were 53% and 47% for the four and six-emotion tasks, respectively. Results indicate that glottal signal features provide good separation of spoken emotions and achieve enhanced classification performance when compared to other approaches.

1. Introduction

Interpersonal communication is greatly facilitated by the detection of emotion through visual and auditory clues, which are used to deduce the motive, intent, and general psychological state of a person. Speech, because of the multilayered processes (cognitive, linguistic, and articulatory) involved in its production, is a main vehicle for emotional expression, which in turn enhances the information contained in the intended spoken message. In intelligent computing, automated recognition of emotion in speech is a growing area of interest with applications that span from speech synthesis and security to psychology, forensic science, health care, and aiding people with disabilities.

In speech communication, phonetic, prosodic, and linguistic features undergo transformations associated with emotional expression. In this context, acoustical analysis aims at the robust extraction of relevant signal features which best describe the changes associated with a particular emotion. A noteworthy notion is that human interaction is carried out over two communication channels: one transmitting the explicit message and the other conveying implicit information about the state of the speakers themselves (see [1], and references therein). Speech analysis provides considerable advantages over other techniques because it is nonintrusive and the signal can be acquired simply with a microphone, even over the telephone, and it has therefore received considerable attention.

The goal of this work is to study the contribution of the glottal flow signal in differentiating emotional states and whether glottal-based features can be effective in spoken emotion recognition.

There are several studies examining speech glottal features under stress (see [2–6], and references therein). Variations of the glottal features have been studied in emotion-related disorders, such as clinical depression in Moore et al. [7, 8] where the spectral tilt and the bias of the glottal frequency response were key features used to derive inter- and intrasentence statistics. Another noteworthy approach using glottal information was undertaken by Ling et al. [9] where voice was considered the output of a Liljencrants-Fant (LF) source model [10] whose glottal formant parameters and spectral tilt can be measured. In that case, glottal frequency characteristics are claimed to be more likely to be preserved in the speech spectrum, as opposed to obtaining the glottal waveform via inverse filtering.

In this work, the effectiveness of using the glottal symmetry, defined as the ratio of closing to opening phase duration, as the key feature in classifying emotion in speech is investigated. Previous work by the authors [11] established that glottal features are rich in conveying emotional information and therefore are quite effective in emotion recognition. Furthermore, emotion classification performance compared favorably against methods using more traditional speech features.

2. Estimation of Glottal Flow Characteristics

During voicing, the air puffs that pass through the glottis when the glottal folds are set in a vibratory mode comprise the glottal flow signal, typically represented as air volume velocity against time. This glottal flow is quasiperiodic and it oscillates at the pitch period, T_0. The complex spectrum of the resulting speech measured at the lips, $S(z)$, can be expressed as:

$$S(z) = G(z)V(z)R(z), \tag{1}$$

where $G(z)$ is the glottal model, $V(z)$ is the vocal tract (VT) system function, and $R(z)$ is the effect of the radiation at the lips [12].

An inverse filtering estimation of the glottal signal can be obtained by solving for $G(z)$ from (1):

$$G(z) = \frac{S(z)}{V(z)R(z)}. \tag{2}$$

The shape of the glottal signal has been extensively investigated and several models specifying its phases from a geometric point of view are available. As summarized by Hardcastle and Laver [13], there are several widely adopted versions, such as those proposed by Rosenberg [14], Hedelin [15], Fant [16–18], Ananthapadmanabha [19], and Ljungqvist and Fujisaki [20]. The model of choice adopted for use in this study is the one proposed by Fant, since it provides better analytical details of the typical shape of a glottal pulse. It is displayed in Figure 1, where U_0 is the peak volume velocity of the glottal pulse, which occurs at t_p, T_o is the opening phase of the pulse, T_c is its closing phase, and FG is defined as the inverse of the glottal pulse width and it signifies the glottal frequency of oscillation. Those parameters can be expressed as:

$$T_c = \left(\frac{1}{\mathrm{FG}}\right)\left[\frac{\cos^{-1}[(k-1)/k]}{2\pi}\right], \qquad T_o = \frac{1}{\mathrm{FG}} - T_c. \tag{3}$$

Glottal symmetry GS is given as the ratio of the closing phase over the opening phase: $\mathrm{GS} = T_c/T_0$.

While intonational features have been traditionally selected for developing systems that classify emotion (see [21], for a review), the role of glottal waveform control in expressing emotional speech has received attention in speech synthesis and voice transformation. Cabral and Oliveira [22] examined the relationship between emotions and glottal parameters and they proposed a system that simulates emotions in neutral speech by changing glottal source parameters and prosody. The role of the glottal amplitude quotient in conveying paralinguistic information in speech was also investigated by Mokhtari and Campbell [23] in the context of automatically annotating large speech databases for use in concatenative speech synthesis.

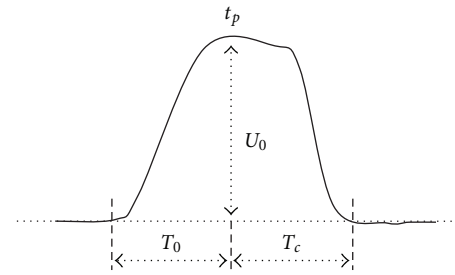


Figure 1: Glottal shape model as proposed by Fant [16].

The glottal flow contribution on the short-time speech spectral envelope is greatly affected by the speaking style. Specifically, while an average of -12 dB/octave roll-off is observed for neutral speech that slope changes to -9 dB/octave for forceful and abrupt speech thus boosting more high frequency energy, and to -15 dB/octave for relaxed speech resulting in a smoother time signal where low frequency energy dominates [24, 25]. A good review of features used in parameterizing the glottal shape is provided by Aims et al. [26].

In this study, the temporal characteristics of the glottal flow were first observed using the EG2-PC laryngograph by Glottal Enterprises. For this purpose, a database was collected from ten male and ten female speakers, each speaking the same 46 sentences in four difference emotions: *angry*, *happy*, *sad*, and *neutral*. Recordings were made at sampling rate of 22050 Hz, 16 bit linear PCM, in two-channels. The first channel was speech recorded with a Neumann TLM103 condenser microphone at 20 cm from the mouth. The second channel was the glottal area function obtained from the laryngograph using neck contacts. The speakers were University of Miami students and not professional actors.

Figure 2 shows an example portion of vowel /E/ from utterance "over there" spoken by a male speaker in the four different emotions. It is noted that the shape of the glottal area waveform is considerably different for speech spoken in different emotion. In our previous work [11], we have studied the effect of selecting appropriate glottal and/or speech features for emotion recognition for a small database and we established that glottal features alone were sufficient. Glottal symmetry, in particular, is the focus of the method described here.

In this work, spoken sentences were short and it was assumed that emotion is clearly conveyed in the first portion of each utterance and it does not change for the remaining sentence or utterance. Therefore, GS values were obtained only from the first five pitch periods of a spoken emotional utterance. Figure 3 shows glottal symmetry smoothed histogram plots for a male and a female speaker as well as for the global average GS value distribution of all speakers, for *angry*, *happy*, *sad* emotions, and *neutral*. GS values were computed

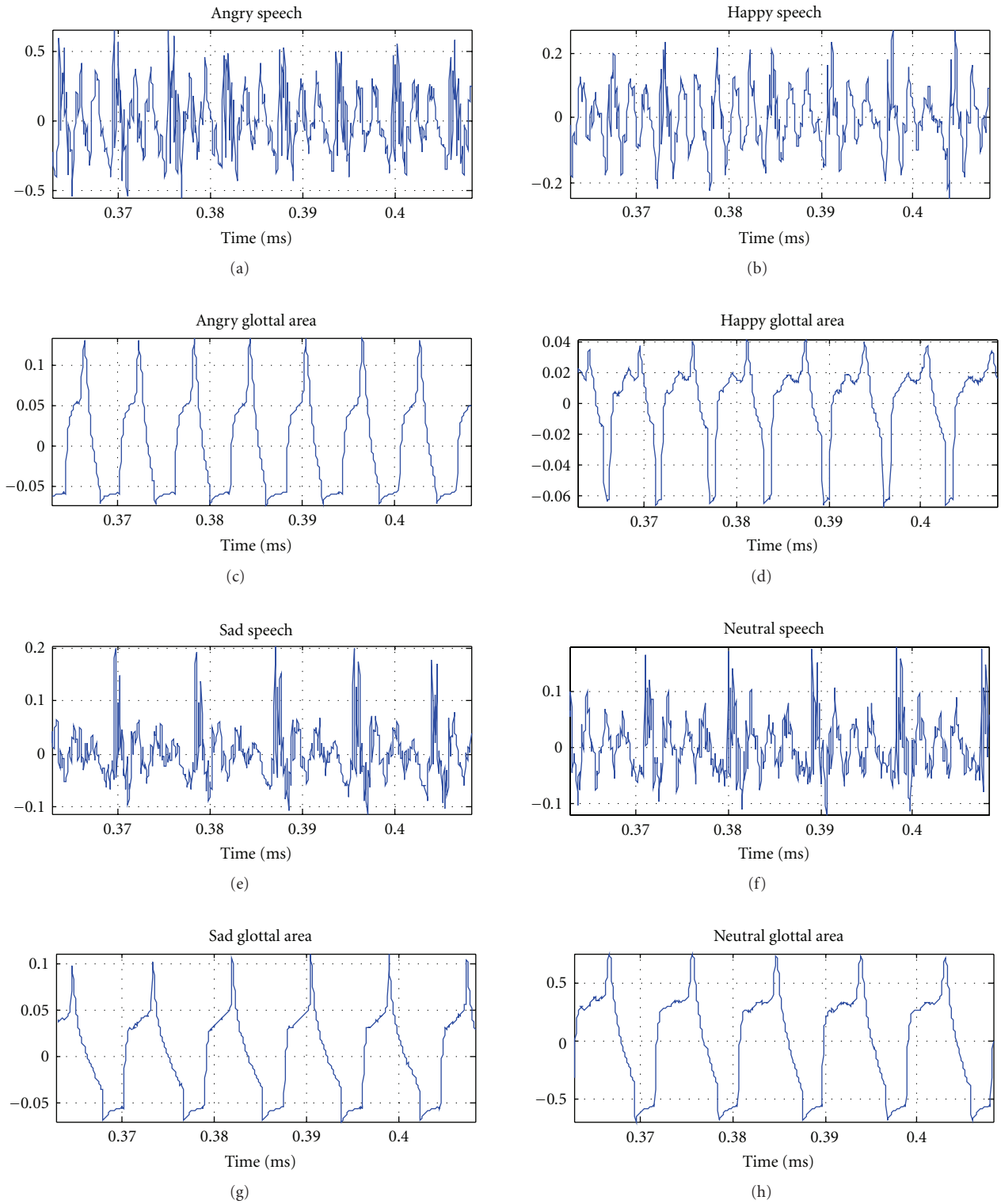


Figure 2: Speech and corresponding laryngograph-obtained glottal area waveforms for emotional speech.

from the glottal flow waveform obtained by inverse filtering, as described in the sequel.

While occasionally in female speakers complete glottal closure may not even occur, it was nevertheless observed that during the production of abrupt and stressed speech the glottal closing phase tended to be shorter, confirming Quatieri's observation that the vocal folds slam faster under stress.

Extracting the glottal flow signal under fluent speech in a practical recording environment can be a challenging task.

However, since the actual durational relationship between the opening and closing glottal phases is important in our task rather than the exact shape of the glottal waveform typical requirements of extracting the exact glottal signal may be relaxed.

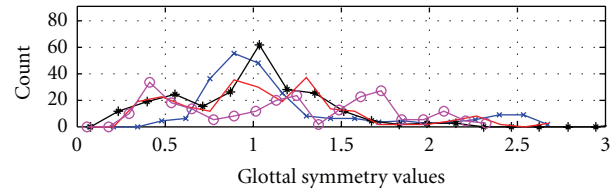
The glottal speed, defined as the inverse of glottal symmetry, was used as one of many speech features in a four-emotion classification task by Yan et al. [27]. However, the effect of a minimal set of features, such as simple temporal measures of the glottal flow, which has been the focus of this work, was not investigated.

The quality of the glottal flow estimation obtained via inverse filtering is critically dependent on accurately approximating the properties of the supraglottal, section of the speech production system. A number of methods dealing with inverse filtering have been proposed [7, 28–30] and they fall into two procedural groups according to the way the volume velocity waveform is obtained: (a) recorded in the mouth as proposed by Rothenberg [28], and (b) recorded outside of (away from) the mouth, thus including the radiation at the lips as in Wong et al. [29]. Quatieri [24] describes a model of the source/vocal tract interaction, which derives the coarse and fine structures of the glottal flow derivative separately. The LF model [10] consisting of seven parameters was used to represent the coarse structure of the flow derivative, while the fine structure (ripple) component was estimated by subtracting the coarse model from the glottal flow derivative obtained via inverse filtering.

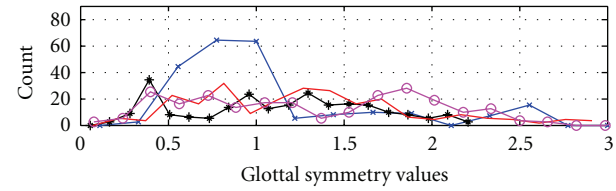
Because of the substantial coupling between the glottis and the dynamically changing supraglottal section the short-time properties of the vocal tract need to be estimated when the glottis is closed. As a result, the fidelity of those methods is conditioned upon the reliable estimation of the glottal opening instants (GOI) and the glottal closure instants (GCI). Generally the closure of the glottis is more abrupt than the opening thus making the identification of GCI much easier. It also places higher precision constraints on the glottal signal extracting algorithm since the associated changes are shorter in time. Analysis is usually done in short window frames of between 10 and 40 ms so that only a few of glottal cycles (usually between 3 and 6) are included. This allows for the estimation of the GCI within 1 to 2 ms of its actual location. Once the GCI location is estimated, the vocal tract model parameters can be determined with better accuracy because the glottis is closed and the signal inside the vocal tract becomes a freely decaying oscillation due to vocal tract resonance.

In order to make the emotion recognition model more indifferent to the alignment between analysis frames and larynx cycles, autocorrelation-based linear prediction was used in the analysis using Durbin's recursion [12, 24]. One major advantage of the autocorrelation method is that its stability is always guaranteed when it is computed in high enough prediction order.

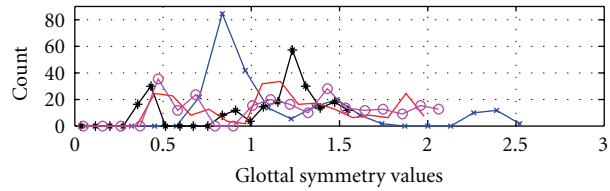
Pre-emphasis filtering typically applied to speech prior to linear prediction introduces a frequency-dependent phase, which is manifested as group delay distortion. This presents a potential problem because GCI location via the inverse filtering depends on accurately reproducing the excitation



(a) A male speaker glottal symmetry ratio values for four emotions



(b) A female speaker glottal symmetry ratio values for four emotions



(c) Average speaker glottal symmetry ratio values for four emotions

Figure 3: Glottal symmetry value distributions for (a) a male speaker, (b) a female speaker, and (c) average for ten male and ten female speakers, for angry (x), happy ($*$), sad ($+$) emotions, and neutral ($\circ$).

peaks in the LP residual. An effective approach is to lower the pre-emphasis filter cutoff frequency as much as possible because the lower that cutoff frequency, the lower the passband group delay.

Many of the issues concerning glottal flow estimation via inverse filtering have been addressed by Brookes et al. [30] including the group delay problem. In essence, the techniques used here compute the frequency-averaged group delay applied to the LP residual signal, while using a sliding window with length of 5 ms. As in all preceding techniques, LP analysis was performed on pre-emphasized speech.

The residual signal was obtained by filtering the speech signal with an all-zero, FIR filter with coefficients provided by the LP analysis. As pointed out by Brookes et al., the use of the LP residual signal requires three assumptions: (a) the VT is assumed to be an all-pole filter, (b) that filter should be estimated solely from the speech waveform, and (c) the LP residual signal will carry the timing instances for identifying the GCI for voiced speech. One of the main characteristics of this method is the addition of an energy-weighted group delay measure to assist in more accurate localization of the impulses, which identify the instances of GCI in the residual signal.

The group delay function of the residual was computed and then averaged across the frequency spectrum. This way, identifying the GCI is more robust as compared to other

inverse filtering techniques. Group delay is defined as the derivative of the phase with respect to frequency, as:

$$\tau_r = \frac{-d\arg(X_r)}{d\omega}, \tag{4}$$

where X_r is the Fourier transform (FT) of a given signal frame $x_r(n)$, and r is the frame location in time. Constant group delay corresponds to linear phase. Addressing the group delay in our analysis can provide better GCI localization within an analysis frame. Expanding (4), the group delay for a sampled signal frame $x_r(n)$, obtained with a window of size N, becomes:

$$\begin{aligned}\tau_r(k) &= -\Im\left(\frac{d\ln(X_r)}{d\omega}\right)\\ &= \Re\left(\frac{\sum_{n=0}^{N-1} nx_r(n)e^{-2j\pi nk/N}}{X_r(k)}\right),\end{aligned} \tag{5}$$

where the numerator is the discrete Fourier transform of the sampled signal $nx_r(n)$ and $\Re$ is real part operator. Noise added to the signal causes the group delay to vary thus affecting the GCI localization. Group delay estimation can be improved by averaging over all discrete frequency bins. Since discrete frequency index, k, assumes integer values, Brookes et al. proposed alternative solutions of how to best estimate the delay.

One option is the average group delay, which is given as:

$$d_{\mathrm{AV}}(r) = \frac{1}{N}\sum_{k-0}^{N-1}\frac{\tilde{X}_r(k)}{X_r(k)}, \tag{6}$$

where $\tilde{X}_r(k)$ is the discrete Fourier transform of $nx_r(n)$. However, the denominator can approach zero for some k resulting in group delay d_{AV} approaching infinity. An alternative group delay measure that addresses this problem, by using the *energy-weighted group delay* or d_{EW}, which limits the bounds of the summation by weighting each term by $|X_r(k)^2|$, the energy at the kth frequency index, is defined as:

$$d_{\mathrm{EW}}(r) = \frac{\sum_{n=0}^{N-1} nx_r^2(n)}{\sum_{n=0}^{N-1} x_r^2(n)}. \tag{7}$$

This measure is bounded within the interval $[0, N-1]$, assuming $x_r(n) \neq 0$. Its relative immunity to noise for SNR $\geq$ 14 dB and its ability to detect the GCI and most other zero crossings makes it appealing. However, when excessive noise is introduced, the center of the energy for the calculated window may shift thus compromising the detection accuracy of GCI.

Another important parameter is the size of the analysis window. If the window is too short (shorter than the length of a glottal period) the captured signal is essentially noise, with zero crossings being detected but not necessarily the GCI. If the window is too large, each impulse at GCI contains a smaller portion of the energy in the frame. This in turn will degrade the time resolution. So there is a tradeoff between time accuracy and detection accuracy. The group delay analysis window length used in this study was 20 ms. Our method was implemented using the dynamic programming projected phase-slope algorithm platform as described by Kounoudes et al. [31], Brookes et al. [30], and Naylor et al. [32].

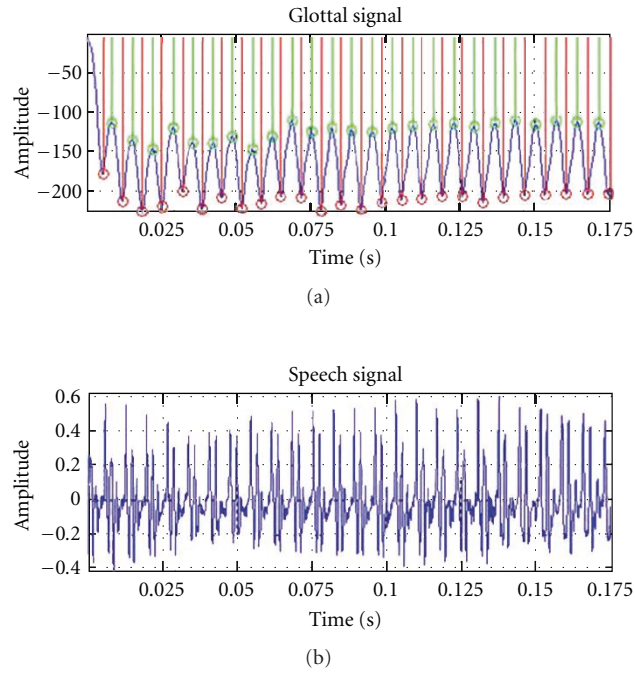


Figure 4: Extracted glottal signal and its corresponding speech signal.

The speech signal was pre-emphasized before processing. The parameters of the vocal tract were estimated in a two-step procedure.

(1) Voiced components were located using the signal envelope and obtained from the magnitude of the Hilbert transform and zero crossing rate. The vocal tract system function, $V(z)$, was estimated during the intervals where the glottis is closed. Because of the small number of available samples during the glottal closed phase, the covariance method was used to estimate the linear prediction model parameters. For a sampling frequency of 8 kHz, a linear prediction order of 8 was selected.

(2) Speech was inverse filtered with the obtained linear prediction model of $V(z)$ to provide the glottal waveform. A laryngograph device was used to simultaneously record the glottal waveform and the obtained signal was subsequently used to verify the results.

Time domain glottal parameters were estimated, such as open quotient OQ $= (T_o + T_c)/T_0$, closing quotient, CQ $= T_c/T_0$ and speed quotient, SQ $= T_o/T_c$. An example of an extracted glottal flow signal that corresponds to a vowel of a neutrally-spoken speech segment is shown in Figure 4, where the locations of the minimum and maximum glottal flow values per cycle are indicated with red and green vertical lines.

Table 1: Emotional utterances available in the Godot corpus.

#	Emotion	Subject 1	Subject 2	Subject 3	Subject 4	Total
1	Angry	141	161	186	0	488
2	Happy	125	48	106	0	279
3	Sad	237	31	127	3	398
4	Neutral	373	181	251	50	855
5	Fear	41	29	55	2	127
6	Surprise	70	40	111	0	221

Table 2: Four emotions recognition rate (%) for balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers.

Actual	Detected			
	Angry	Happy	Sad	Neutral
Angry	**98.74**	0.51	0.42	0.33
Happy	24.11	**75.47**	0.18	0.24
Sad	35.24	0.13	**64.59**	0.04
Neutral	0.38	20.56	15.18	**63.88**

3. Speech Corpus Design

In contrast to speech recognition research where many databases are available that allow benchmarking algorithm performance for a variety of applications, availability of spoken emotion corpora is still limited in comparison. Those databases that are available differ greatly in their size, number of speakers, number of emotions, recording setup, type of speech (natural or acted, vocabulary type and size), target application (recognition or synthesis), and spoken language. Ververidis and Kotropoulos [21] summarized a total of 64 emotional speech data collections which included a total of 29 different emotions. Out of those, only four corpora included six emotional classes, as needed in the current study, of which only one was in English and that was very limited since it contained recordings from one speaker only.

Creating a collection containing multispeaker emotional recognition speech, which includes a good variety of emotions, in the language of choice and custom-selected recording conditions, is quite challenging and costly. The "Emotional Prosody Speech and Transcripts" dataset for example available at the Linguistic Data Consortium was a close match for the needs of this research. It still however provided very limited choice of speech examples, since the recorded utterances were spoken dates and numbers only. On the other hand, the "Berlin Database" was a better match containing six emotions and good number of utterances, but it was recorded in German, thus remained out of the scope of this research.

For this work, an emotional speech database was developed based on an audio recording of the theatrical play "Waiting for Godot", written by Samuel Beckett in 1949. The recording was released on April 3, 1961, it involves four speakers and it is approximately 100 minutes long. This play has been voted by experts as "the most significant English language play of the 20th century" [33]. The characters, the actors that portrayed them, and their age were: *Subject 1*: Gogo/Estragon (Zero Mostel) age 46; *Subject 2*: Pozzo (Kurt Kasznar) age 47; *Subject 3*: Didi/Vladimir (Burgess Meredith) age 53; *Subject 4*: Child (Luke Halpin) age 13. The test sets were split in two main groups: tests on individual speakers and tests on combined speaker sets.

Manual labeling for six different emotional classes was performed on the original speech. These were: *happy*, *angry*, *sad*, *fear*, *surprise*, and *neutral*. The applied methodology used two human labelers who segmented the speech signal into the target emotional classes, marking the beginning and end of each emotion using a signal marking utility. This way, one emotion was followed in time by another emotion. The two human labelers were instructed to label all speech according to the stated six emotions. In a second phase, labelers were informed of the agreed labels and were asked to reconsider the labels where the other party differed, without knowing the nature of the disagreement. Finally, labeling disagreements were resolved in a joint, open session to provide one, final, labeled database.

The speech corpus consists of three male and one child speakers, speaking in random order, and it contains a total of 2,368 emotional utterances (turns). Their distribution is displayed in Table 1. The small portion of the child's speech mainly contained *neutral* emotion (*Subject 4*) and therefore it was not used in subsequent work.

The play was originally recorded in analog and included audible background, broadband noise levels. Speech was provided at sampling frequency of 22050 Hz in a single channel with 16 bits per sample linear quantization. For convenience and without loss of important information, the corpus was downsampled to 8 kHz. In order to remove biases and to balance the corpus, 127 utterances from each emotion were randomly selected to match the smallest emotion population, which corresponded to *fear*. The included utterance lengths varied in time between one and six seconds.

Table 3: Six emotions recognition rate (%) for a balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers.

Actual	Detected					
	Angry	Happy	Sad	Neutral	Fear	Surprise
Angry	**85.37**	12.18	0.44	1.56	0.18	0.27
Happy	30.44	**67.32**	0.09	1.44	0.44	0.27
Sad	26.96	5.91	**66.02**	0.49	0.42	0.20
Neutral	15.67	4.04	23.44	**56.51**	0.18	0.16
Fear	22.31	8.64	10.31	1.29	**57.07**	0.38
Surprise	20.27	4.58	7.49	0.56	0.33	**66.77**

Table 4: Four emotions recognition rate (%) for a balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers after LPF.

Actual	Detected			
	Angry	Happy	Sad	Neutral
Angry	**51.62**	43.70	2.35	2.33
Happy	0.56	**93.54**	3.14	2.76
Sad	26.35	25.34	**46.60**	1.71
Neutral	10.19	23.95	16.77	**49.09**

For the development of the classification systems, each set was randomly split in two parts: 80% used for training and 20% for testing.

Depending on the viewpoint, there were a series of different category tests performed on the corpus.

(1) From signal quality standpoint, tests included:

(i) originally-recorded speech;
(ii) noisy speech with SNR of 10 dB and 30 dB;
(iii) low-pass filtered (LPF) speech, with an FIR, linear phase filter with passband from 0 to 600 Hz and stopband starting at 800 Hz and 80 dB attenuation, which imposed severe degradation of the speech to the point of making it incomprehensible.

(2) From data presenting and arrangement point of view, the tests included were:

(i) utterance-based balanced, speaker and text independent with 100 utterances per emotion, and random number of speakers;
(ii) glottal symmetry-based balanced, speaker and text independent both per emotion and per speaker.

The computed values of the glottal symmetry of five consecutive glottal periods detected in the first voiced segment of each speaking actor's turn were used to create a 5-dimensional feature vector for each emotional utterance.

4. GMM Deployment

The Gaussian Mixture Model (GMM) has been used extensively in many engineering applications in which data can be viewed as generated from multiple mixed sources of a certain type of distribution, based on a set of corresponding prior probabilities. Besides its inherent modeling power of fitting probability densities arbitrarily [34], it is particularly suited for statistical pattern classification by the use of the Expectation-Maximization (EM) procedure [35] for GMM parameter estimation. There have been many efforts reported using GMM in emotion recognition with good results as regards the recognition rate. However, there are still fundamental issues of importance in the employment of GMM for emotion recognition, like for other GMM applications such as: model initialization, determination of the number of components, the estimation of prior probabilities of classes, normalization of feature vectors, and handling of singularity issue. In this work, in addition to recognition performance, the effect of using a different number of components, the normalization of feature vectors, and ways to handle matrix singularity in EM computation, were investigated. These superparameters can make a significant impact on the computation load, convergence, and system performance.

In this work, all GMMs for each emotional state have diagonal covariance matrices instead of full covariance. The number of components used in GMM regarding the different system configurations related to the different dimensionalities of feature vectors and the size of training samples. To handle overflow in computing the covariance matrices and their inverse matrices as well, appropriate techniques were employed, including variance flooring [36], relative variance flooring [37], and appropriate normalization of the features. In this work, eight Gaussian mixtures were adopted as they proved to be the most effective.

5. Emotion Classification Results

Two emotional speech sets were considered: a six-emotion set that included *angry*, *happy*, *sad*, *neutral*, *fear*, *and surprise*, and a four-emotion set containing *angry, happy, sad, and neutral,* which was created by removing the speech portions

Table 5: Six emotions recognition rate (%) for a balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers after LPF.

Actual	Detected					
	Angry	Happy	Sad	Neutral	Fear	Surprise
Angry	**44.64**	28.76	5.21	15.28	6.11	0.00
Happy	7.35	**64.38**	4.83	15.85	7.59	0.00
Sad	16.99	24.66	**39.42**	13.18	5.75	0.00
Neutral	10.06	19.23	8.03	**57.38**	5.30	0.00
Fear	9.27	26.07	5.64	7.61	**51.41**	0.00
Surprise	8.76	16.56	6.22	6.60	9.23	**52.63**

Table 6: Four emotions recognition rate (%) balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 30 dB.

Actual	Detected			
	Angry	Happy	Sad	Neutral
Angry	**93.91**	0.41	3.51	2.17
Happy	39.63	**54.70**	3.05	2.62
Sad	26.52	0.71	**70.51**	2.26
Neutral	0.75	23.74	23.70	**51.81**

that included *fear* and *surprise*, as if they were never produced.

Principal component analysis (PCA) on the first and second glottal pulses encountered at the beginning of a voiced emotional speech segment was performed to provide a representation of the separability for the six-emotion classification problem and it is graphically shown in Figure 5. It is apparent that although classes have a substantial overlap in the middle of the plot, partial separation of each emotional class by just examining two neighboring glottal pulses in a sequence is possible due to the clusters formed away from the center.

Emotion recognition performance was tested using clean speech, severely filtered speech and noisy speech. The two distortion conditions were applied prior to glottal flow extraction to test the method's performance in real-world conditions.

5.1. Classification Performance for Clean Speech. *Angry*, *happy*, *sad*, and *neutral* were included in a four-emotion subset. As shown in Table 2, performance for a balanced 100-utterances per emotion set ranged from 64% for *neutral* to 98.7% for *angry*, indicating that *angry* is relatively quite a distinct emotion. All emotions were quite distinct from *neutral*, although *neutral* had substantial confusion with *happy* and *sad*, and *sad* was confused for *angry* in about one out of three cases.

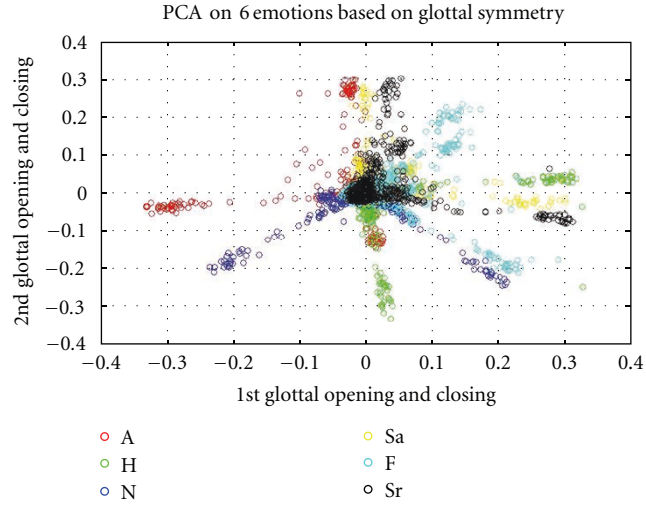


Figure 5: PCA analysis of the 1st versus 2nd Glottal Symmetry for 6 emotions.

Recognition of all six emotions (*angry*, *happy*, *sad*, *neutral*, *fear*, *and surprise*) was tested on a balanced 100-utterances per emotion set as well. Results are in Table 3, indicating that *angry* is still relatively quite a distinct emotion with the highest recognition rate of 85.4%, *neutral* having the lowest rate of 56.5%, while *sad* and *happy* are most confused with *angry*.

As expected, average four-emotion classification is better for the four-emotion system with recognition performance at 75.7%, while for the six-emotion system performance was at 66.5%.

5.2. Classification Performance for Low-Pass Filtered Speech. All clean speech was passed through a low-pass filter with cutoff frequency at 600 Hz, resulting in severely distorted and unintelligible speech. Classification results are shown in Tables 4 and 5. As expected, performance was substantially lower than in the case of clean speech, but still useful. However, there were important differences in those results as compared to clean speech. In both the four- and the

Table 7: Six emotions recognition rate (%) balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 30 dB.

Actual	Detected					
	Angry	Happy	Sad	Neutral	Fear	Surprise
Angry	**80.77**	0.90	12.84	2.28	1.40	1.81
Happy	37.20	**45.38**	11.18	2.86	1.66	1.72
Sad	24.09	0.58	**70.07**	2.45	1.46	1.35
Neutral	22.92	0.54	27.48	**47.10**	1.08	0.88
Fear	30.26	0.45	30.92	2.30	**33.96**	2.11
Surprise	28.37	0.47	27.27	2.39	1.33	**40.17**

Table 8: Four emotions recognition rate (%) balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 10 dB.

Actual	Detected			
	Angry	Happy	Sad	Neutral
Angry	**50.96**	42.09	4.28	2.67
Happy	34.01	**58.54**	4.52	2.93
Sad	22.45	29.35	**45.34**	2.86
Neutral	14.01	28.27	7.40	**50.32**

six-emotion tests, *happy* was the emotion recognized most successfully. This correlated well with the fact that *happy* is typically produced with more relaxed glottal operation that results in higher frequencies having lower spectral energy and information, as opposed to *angry*, which is typically produced with abrupt glottal closure and therefore results in a signal with more features at higher frequencies. If high frequency information is removed by severe low-pass filtering, then *happy*, the emotion less dependent on high frequency features would be the least affected. In the four-emotion tests, *happy* achieved the best recognition rate of 93.5%, while *sad* had the lowest score of 46.6%. In the six-emotion test *happy* achieved the best recognition rate of 64.4% and *sad* had the lowest score of 39.4%. As expected, four-emotion classification results are higher with an average performance of 60.2%. The six-emotion classifier achieved average recognition performance of 51.6%.

The performance results included in Tables 4 and 5 (which is quite good considering the dramatic filtering) suggests the importance of low-frequency information in emotion recognition and helps to reinforce the case of glottal features.

5.3. Classification Performance for Additive White Gaussian Noise at SNR = 30 dB and SNR = 10 dB. In addition to the tests performed in clean and LPF speech, white Gaussian noise was added to clean speech at SNRs of 30 and 10 dB before the glottal symmetry was estimated. Test results for 30 dB are displayed in Tables 6 and 7; those for 10 dB are shown in Tables 8 and 9.

For the four-emotion test at 30 dB, shown in Table 6, the results were still high in comparison to clean speech conditions (Table 2), ranging from 94% for *angry* down to 51.8% for *neutral*, with average performance at 68%. The trend was similar for the six-emotion test at 30 dB, with *angry* recognized 80.8% of the time, while *fear* had the lowest performance at 34%, as shown in Table 7, with average performance at 53%.

In the 10 dB test, recognition performance was lower but still useful. For the 4-emotions test, *happy* was recognized most successfully at 58.5% of the time, while *sad* had the lowest score, with average classification performance at 53%. In contrast, for the six-emotion test, *fear* was the emotion recognized best, while *angry* had the lowest recognition rate (37.1%) and the average classification performance reached 47%.

Finally, the statistical significance of our results was estimated [38] and indicated that indeed our results are significant, as summarized in Table 10.

6. Conclusions

In this work, we examined the effectiveness of using features of the glottal airflow during voicing for emotional speech classification. The relationship between the glottal signal dynamics and the expression of stressed or emotional speech has been examined in other past studies. Other authors have also completed comparison of speech and glottal features in their effectiveness for emotion recognition using different classifiers. In this paper, we have shown that glottal symmetry is a simple but quite effective speech feature which can provide high classification performance for spoken emotions.

For this purpose, we developed and used a large, multispeaker database based on an audio recording of Samuel Beckett's "Waiting for Godot". Four- and six-emotion classification tasks were pursued under clean and distorted

Table 9: Six emotions recognition rate (%) balanced speaker and text independent test-utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 10 dB.

Actual	Detected					
	Angry	Happy	Sad	Neutral	Fear	Surprise
Angry	**37.14**	18.87	7.79	14.28	20.84	1.08
Happy	16.20	**40.79**	8.65	12.96	20.58	0.82
Sad	12.79	12.04	**43.66**	13.03	17.45	1.03
Neutral	2.12	8.82	18.70	**50.84**	18.29	1.23
Fear	8.05	12.04	7.98	11.61	**59.29**	1.03
Surprise	3.75	9.47	10.02	8.44	19.21	**49.11**

Table 10: Statistical significance for four and six emotions using glottal symmetry on 100 utterances of clean speech.

Number of emotions	Degrees of freedom	Chi-square	Statistical significance
4	3	31.62	>99.99
6	5	72.75	>99.99

conditions that included noisy and severely filtered signal. Averaged emotion recognition results were at 75% in the four-emotion test and 62% in the six-emotion test in clean signal conditions. Average performance was 60.8% and 54.3%, respectively, for low-pass filtered speech, and it ranged from 68% to 53% for the SNR = 30 dB condition, to 53% to 47% for the SNR = 10 dB condition, for each of the four- and six-emotion recognition tasks, respectively. Our results are statistically significant.

In future work, separation of utterances in different emotional classes when building the corpora should be verified through listening tests performed by multiple subjects. This will give more clarity on how the listeners relate to a particular dataset and it will help mitigate any bias connected to either speech quality or perceptual cognition that may negatively impact the correct separation of emotions when forming the corpora.

For more practical applications, since every emotion may carry different levels of intensity, from *weak to strong*, the degree of expression, each emotion may also be tested.

Finally, in this work when an emotional utterance was evaluated only a small voiced fragment from the beginning of the utterance was taken for analysis, which may or may not hold the most effective information for the emotion sought. A more sophisticated approach for signal frame selection and processing may produce better or more robust results. For on-line systems, the temporal variations of features may be studied using Hidden Markov Models (HMM). Finally, fusion of acoustic and linguistic features may also be considered for achieving more robust emotion recognition.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] K. E. Cummings and M. A. Clements, "Analysis of glottal waveforms across stress styles," in *Proceedings of the IEEE International Conference in Acoustics, Speech, and Signal Processing*, pp. 369–372, April 1990.

[3] K. Cummings and M. Clements, "Improvements to and applications of analysis of stressed speech using glottal waveforms," in *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, pp. 25–28, 1992.

[4] K. E. Cummings and M. A. Clements, "Application of the analysis of glottal excitation of stressed speech to speaking style modification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 207–210, April 1993.

[5] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 88–98, 1995.

[6] A. M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "Physical variations related to stress and emotional state: a preliminary study," *Journal of Phonetics*, vol. 24, no. 3, pp. 313–335, 1996.

[7] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2849–2852, September 2003.

[8] E. Moore, 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 Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, 2008.

[9] Z. H. Ling, Y. Hu, and R. H. Wang, "A novel source analysis method by matching spectral characters of LF model with STRAIGHT spectrum," in *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII '05)*, vol. 3784, pp. 441–448, 2005.

[10] G. Fant, "Glottal flow: models and interaction," *Journal of Phonetics*, vol. 14, pp. 393–399, 1986.
[11] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech and Language*, vol. 24, no. 3, pp. 445–460, 2010.
[12] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.
[13] W. Hardcastle and J. Laver, *The Handbook of Phonetic Sciences*, Blackwell Publishers Ltd, 1999.
[14] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America*, vol. 49, pp. 583–590, 1971.
[15] P. Hedelin, "A glottal LPC-vocoder," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84)*, pp. 1.6.1–1.6.4, San Diego, Calif, USA, 1984.
[16] G. Fant, "Glottal source and excitation analysis," *Speech Transmission Laboratory, Quarterly Progress and Status Report*, no. 1, pp. 70–85, 1979.
[17] G. Fant, "The voice source—acoustic modeling," *Speech Transmission Laboratory, Quarterly Progress and Status Reports*, no. 4, pp. 28–48, 1982.
[18] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory, Quarterly Progress and Status Report*, no. 4, pp. 1–13, 1985.
[19] T. V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *Speech Transmission Laboratory, Quarterly Progress and Status Report*, vol. 25, no. 2-3, pp. 1–24, 1984.
[20] M. Ljungqvist and H. Fujisaki, "A method for simultaneous estimation of voice source and vocal tract parameters based on linear predictive analysis," *Transactions of the Committee on Speech Research, Acoustical Society of Japan*, vol. S85-21, pp. 153–160, 1985.
[21] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.
[22] J. P. Cabral and L. C. Oliveira, "Emovoice: a system to generate emotions in speech," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, pp. 1798–1801, Pittsburg, Pa, USA, 2006.
[23] P. Mokhtari and N. Campbell, "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech," *IEICE Transactions on Information and Systems*, vol. E86-D, no. 3, pp. 574–582, 2003.
[24] T. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*, Prentice Hall, 2002.
[25] J. M. Picket, *The Sounds of Speech Communication*, Pro-Ed, Inc., Austin, Tex, USA, 1980.
[26] M. Aims, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 2145–2148, September 2005.
[27] Z. Yan, Z. Li, Z. Cairong, Y. Yinhua, H. Chengwei, and W. Qingyun, "Modified quadratic discrimination function for non-normal distribution and its application in speech emotion recognition," in *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS '08)*, pp. 213–216, December 2008.
[28] M. Rothenberg, "A new inverse filtering technique for deriving the glottal air flow waveform during voicing," *Journal of the Acoustical Society of America*, vol. 53, pp. 1632–1645, 1973.
[29] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.
[30] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 456–465, 2006.
[31] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proceeding of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 349–352, Orlando, Fla, USA, May 2002.
[32] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.
[33] N. Berlin, "Traffic of our stage: why waiting for Godot?" in *The Massachusetts Review*, 1999.
[34] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.
[35] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–39, 1977.
[36] F. Bimbot, M. Blomberg, L. Boves et al., "An overview of the CAVE project research activities in speaker verification," *Speech Communication*, vol. 31, no. 2, pp. 155–180, 2000.
[37] Y. Zhang and M. Scordilis, "Optimization of GMM training for speaker verification," in *Proceedings of the IEEE Speaker and Language Recognition Workshop (ODYSSEY '04)*, pp. 236–231, Toledo, Spain, May-June 2004.
[38] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.