



On the Importance of Glottal Flow Spectral Energy for the Recognition of Emotions in Speech

Ling He¹, Margaret Lech¹ and Nicholas Allen²

¹School of Electrical and Computer Engineering, RMIT, Melbourne, Australia

²Department of Psychology, University of Melbourne, Melbourne, Australia

{ling.he, margaret.lech}@rmit.edu.au, nba@unimelb.edu.au

Abstract

Two new approaches to feature extraction for automatic emotion classification in speech are described and tested. The methods are based on recent laryngological experiments testing the glottal air flow during phonation. The proposed approach calculates the area under the spectral energy envelope of the speech signal (AUSEES) and the glottal waveform (AUSEEG). The new methods provided very high recognition rates for seven emotions (contempt, angry, anxious, dysphoric, pleasant, neutral and happy). The speech data included 170 adult speakers (95 female and 75 male). The classification results showed that the new features provided significantly higher classification results (89.95% for AUSEEG, 76.07% for AUSEES) compared to the baseline MFCC approach (37.81%). The glottal waveform based AUSEEG features provided better results than the speech based AUSEES features, indicating that the majority of the emotion information is likely to be added to speech during the glottal wave formation.

Index Terms: nonlinear modeling, emotion classification, speech analysis, glottal energy, feature extraction.

1. Introduction

Speech is the most important means of communication among humans. Apart from conveying linguistic information between speakers, it also carries a large paralinguistic content including vital information about speakers' emotions, personalities, attitudes, feelings, levels of stress and current mental states. As a biological signal, speech contains a lot of medical diagnostic information and psychological behavioral information, which in comparison with other biological signals, such as for example ecg or eeg, has been very much under-utilized. One of the reasons for this under-utilization of the vital information present in speech is the combination of high complexity and a wide bandwidth of the speech signal, which makes the analysis relatively more complex than in the case of other bio-signals. Another limiting factor is a serious lack of proper modeling and understanding of the speech production process. The classical source-filter model has a linear character and was generated a few decades ago for the purposes of telecommunication engineering, where conveying of an accurate linguistic content was of primarily importance. It does not include mechanisms explicitly responsible for the generation of the paralinguistic aspect of speech. As a result the majority of the current approaches to emotional speech analysis rely on the assumption that the emotional state of a speaker affects in some way speech parameters assumed by the existing source-filter model. Subsequently these parameters

including the fundamental frequency F_0 , formants and energy, or parameters derived from them, are the most often cited in the literature as characteristic features used in emotion recognition from speech [8,9]. An increasing number of recent laryngological and psychological studies aim to improve our understanding of mechanisms involved in speech production and in particular the generation of the paralinguistic aspects of speech [11,12].

This study follows the results of recent laryngological experiments [3,10] investigating the nonlinear characteristics of air flow during the phonation process. Based on the suggestions presented in these reports, two new types of characteristic features are proposed and applied to emotion recognition in speech. It is demonstrated that the proposed features provide significantly better performance than the conventional emotion recognition method based on the mel frequency cepstral coefficients (MFCC). Moreover, the results based on the new features indicated that the emotional aspect of speech is likely to be generated mostly during the glottal flow formation, and the spectral distribution of the glottal energy is an important factor for differentiating between emotions.

Importantly, the speech data used in this study includes natural speech produced within family environments during typical conversations between family members.

The remaining sections of this paper are organized as follows. Section 2 explains the nonlinear model of the glottal flow formation resulting from a number of recent studies. Section 3 describes the speech data base used in this study. Section 4 explains the methodology of emotion classification and introduces two new types of characteristic features. Section 5 describes the experiments and results, and Section 6 presents the conclusions.

2. Nonlinear Model of the Glottal Flow Formation

The classical source-filter theory of voice production assumes the air flow through the vocal folds and vocal tract has a unidirectional and laminar character. However, recent advances in theoretical acoustics and computational modeling, and experiments in mechanical models suggest that certain modifications are needed.

Teager [4] indicated the existence of nonlinear air vortices formed during speech production. These findings were later confirmed through experimental studies of fluid flow in a dynamic mechanical model of the vocal folds and tract [5].

In a study of speech classification under stress Zhou *et al.*, [2] proposed that in the emotional state of anger or stress, additional sound sources can be generated in the form of air vortices. Features sensitive to the presence of these additional vortices can indicate the emotional state of a speaker. In recent

experimental studies [3,10] using excised canine larynges, two types of consistent, periodic air vortices were identified. During the early opening phase of the vocal folds, when the glottis is convergent, *supraglottal vortices* occur above the vocal folds. During the latter part of the vocal fold closing, when the glottis is divergent, *intraglottal vortices* are formed between the vocal folds. The *intraglottal vortices* generated between symmetrically vibrating vocal folds produced a negative pressure, resulting in a suction force promoting a rapid closing of the vocal folds. The acoustic consequence of the rapid flow shutoff is an increase of energy in the higher harmonics when compared to asymmetrically vibrating vocal folds. The *supraglottal vortices* on the other hand provide additional sound sources when hitting hard surfaces of the vocal tract or interacting with each other. The presence of these additional sound sources is manifested as additional harmonics and cross-harmonics in the speech spectra. Figure 1 shows a new nonlinear model of the air flow during the phonation process based on findings reported in [2] and [3,10]. Note that the supraglottal vortices do not necessarily coincide in time with the intraglottal vortices.

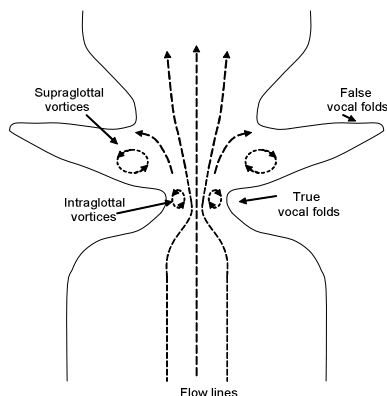


Figure 1: *New nonlinear model of the glottal flow formation based on [2] and [3,10].*

In Khosla [3], a microphone was placed 8 inches downstream of the vocal folds to record the sound. The results showed a strong correlation between the degree of symmetry of the vocal folds vibration and the distribution of the acoustic energy across frequencies. A relative increase of energy of the high frequency harmonics was observed in the case of non-symmetric vocal folds when compared to the symmetric vocal folds. Assuming that the same holds true for humans, it can be hypothesized that an emotional state of a speaker can alter the viscosity and elasticity of the vocal folds vibration providing a shift in the spectral energy distribution of the glottal waveform.

Aiming to test the above hypothesis, new types of features were proposed. The first type, calculates the area under the spectral energy envelope of the speech signal (AUSEES) and the second method calculates the area under the spectral energy envelope of the glottal waveform (AUSEEG).

It was assumed that if our hypothesis is true, these two methods would show a relatively high emotion classification rates in speech. In addition, a comparison between the correct classification rates for AUSEES and AUSEEG would provide an indication whether the emotional aspect is added to speech during the glottal flow formation or during the vocal tract articulation (filtering) process.

3. Speech Data

The speech data was obtained as a result of research cooperation with the Oregon Research Institute (ORI) in USA. The data set consisted of video recordings collected for the purpose of behavioral studies within family environments. The sound track of the video recordings contained typical conversations between family members (parents and children) during problem solving interactions. The data selected for this study contained speech of 170 adult speakers (parents) including 95 female speakers and 75 male speakers. The speakers had a general USA-English accent. The speech recordings were down sampled from the original sampling rate of 44.1 kHz to 22 kHz. The videotapes were annotated in real time by trained psychologist using the Living in Family Environments (LIFE) coding system [1]. The annotation provided second-by-second labeling of the following seven emotions: contempt, angry, anxious, dysphoric, pleasant, neutral and happy. Each emotion was represented by a number of speech utterances of an average duration of 1.5 seconds. The numbers of recordings representing each emotion are listed in Table 1.

Table 1. *Description of the speech database.*

	Number of recordings (each of approx. length 1.5 sec)						
	contempt	anger	anxious	dysphoric	pleasant	neutral	happy
Females	111	145	146	143	153	142	143
Males	92	146	142	144	144	146	143
Both	203	291	288	287	297	288	286

4. Method

In the pre-processing stage the amplitudes of speech samples were normalized into the range $<-1;+1>$. After removal of noise, and voiced/silence detection, the voiced speech frames were concatenated and used in the two-stage processing illustrated in Figure 2. In the first stage (training) characteristic features representing known emotions were used to train the emotional class models. In the second stage (classification or testing), characteristic features from speech samples of unknown classes were compared with the models to determine emotional classes to which they belonged.

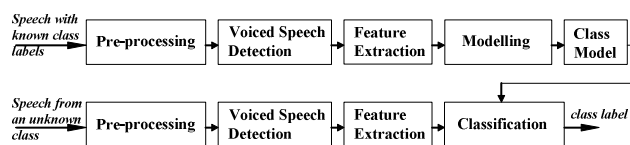


Figure 2: *Flowchart of the training and classification process.*

4.1. Feature Extraction

4.1.1. Conventional MFCC Features

The mel-frequency cepstral coefficients (MFCCs) are widely used acoustic features for speech modeling and pattern recognition [7,8]. For each speech frame, the Fourier transform and the energy spectrum were calculated. The energy spectrum was then mapped onto the mel-frequency scale. Finally, the discrete cosine transform (DCT) of the mel log powers was calculated and the first 12 DCT coefficients provided the MFCC values corresponding to a given frame.

The MFCC parameters were used as a benchmark for comparison with the proposed features AUSEEG and AUSEES.

4.1.2. New AUSEEG and AUSEES Features

The classical source-filter theory of voice production assumes that the air flow through the vocal folds (source) and the vocal tract (filter) is unidirectional. During phonation, the vocal folds vibrate. One vibration cycle includes the opening and closing phases in which the vocal folds are moving apart or together, respectively. The number of cycles per second determines the frequency of the vibration, which is subjectively perceived as pitch or objectively measured as the fundamental frequency F0. The sound is then modulated by the vocal tract configuration and the resonant frequencies of the vocal tract are known as formants. The calculation of glottal flow characteristics provides a lot of challenges. The primary problem lies in the difficulty of separating glottal and vocal tract characteristics in the acoustic speech waveform. On the other hand, as indicated in [6], glottal features appear to be strongly correlated with depression, which indicates that they could be also useful in stress and emotion classification.

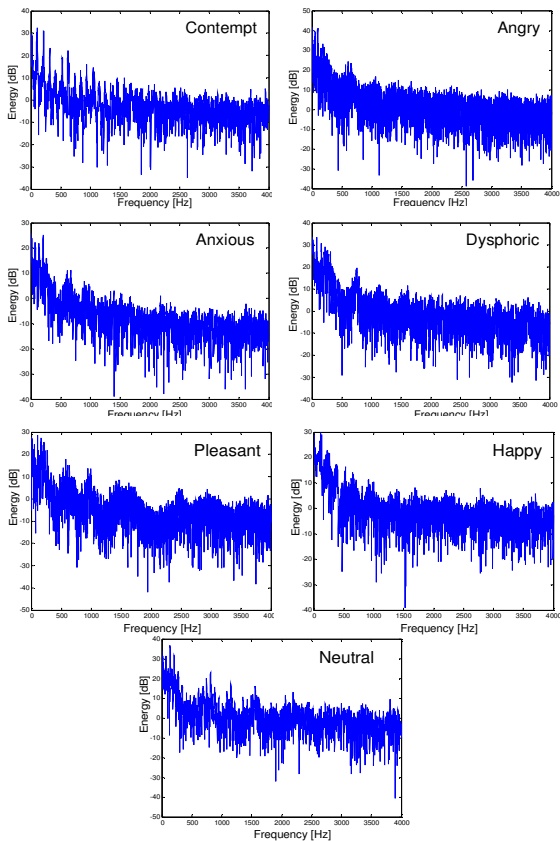


Figure 3: Examples of the spectral energy envelopes of glottal waveforms produced with different emotions.

In Figure 3, glottal spectra for the speech signal recorded under different emotions are plotted. It can be seen that different emotions have different amplitude gradients and distributions of energy across frequency. The examples shown in Figure 3 were for speech utterances of approximately 1.5 seconds duration and expressed with different emotions. It should be noted that these examples were calculated for utterances with different linguistic content. Since the speech data represented natural speech it was not practical to search

for identical utterances pronounced with different emotions. Despite of this limitation, Figure 3 provides two important observations. Firstly, the spectral energy decreases with frequency, however the rate of this decrease differs across emotions. Secondly, the numbers and values of spectral peaks representing different harmonics also vary across emotions. The slopes of the glottal spectral envelope were tested as a possible feature for emotion classification. A line was fit to the first 8 harmonics representing a speech frame; however the results were relatively poor. This was partially due to difficulties associated with determining frequencies of the harmonic components. It was found that the value of the area under the spectral envelope provided much more stable parameters. Based on this, two closely related types of features AUSEEG and AUSEES were proposed. The calculation steps for the AUSEEG and AUSEES methods are illustrated in Figure 4. In both cases, the FFT algorithm was applied either to the glottal waveform (AUSEEG) or to the voiced speech signal (AUSEES), and the spectral energy was calculated for each frame. The energy levels below an arbitrary threshold $\zeta = 0\text{dB}$ were set to zero. The value of ζ was determined experimentally. The entire spectral range of 11 kHz was subdivided into 16 spectral sub-bands of equal width of 687.5 Hz on a linear scale. For each sub-band the area under the spectral envelope was calculated generating 16 feature parameters representing a given frame. It was determined through classification tests that the linear equidistant frequency sub-bands provided better classification results than the logarithmically equidistant sub-bands. As indicated in [15] characteristic features which provide a high resolution at both low and high frequency bands are essential in stress and emotion recognition in speech. In the case of AUSEEG, the glottal waveform was estimated using the inverse filtering and integrator [14].

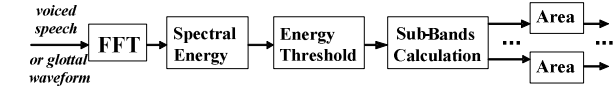


Figure 4: AUSEEG/AUSEES calculation steps.

4.2. Classification

The two most frequently used classifiers were employed to generate emotional models and test the speech samples. These were the Gaussian mixture model (GMM) and the k-nearest neighbors (KNN) method [13]. The classification score for all classifiers was calculated as an average percentage of identification accuracy (APIA) defined as follows:

$$APIA = \frac{1}{N_r} \frac{N_c}{N_T} \times 100\% \quad (1)$$

Where N_c is the number of test inputs correctly identified, N_T is the total number of test inputs, and N_r is the number of repeated runs.

5. Experiments and Results

The emotion classification process was performed using three different feature extraction methods: AUSEEG, AUSEES and MFCC. Each feature extraction method was tested using two different classifiers: GMM and KNN. In the tests involving the GMM classifier 5 Gaussian mixtures were used to model each emotional class. For each feature/classifier combination, the training and classification process was run 15 times, each time with different randomly chosen training and testing sets. Eighty percent of the entire data was used in the training

process and 20% in the testing. The classification results were assessed using the average percentage of identification accuracy given in Eq.(1). The emotion recognition was tested for each gender separately, as well as for the entire data including both genders. In all cases, the speech samples were classified into seven classes of emotion: contempt, angry, anxious, dysphoric, pleasant, neutral and happy. Table 2 provides a performance comparison based on the average percentage of identification accuracy for all feature/classifier combinations, for two genders and for entire data including both genders. In Tables 3 and 4, the average percentage of identification accuracy for each emotion using the new feature extraction methods AUSEEG (Table 3) and AUSEES (Table 4) are presented.

Table 2. The average percentage of identification accuracy using three types of features: AUSEEG, AUSEES and MFCC.

Datasets	AUSEEG		AUSEES		MFCC	
	GMM	KNN	GMM	KNN	GMM	KNN
Females	85.17	79.25	81.70	69.80	45.20	39.29
Males	84.79	84.86	78.96	67.22	43.54	38.65
Both	89.95	84.27	76.07	71.32	37.81	32.21

Table 3. The average percentage of identification accuracy for each emotion using AUSEEG features and GMM.

Group	Emotions						
	Contempt	Angry	Anxious	Dysphoric	Pleasant	Neutral	Happy
Females	98.79	97.14	94.67	63.81	96	75.24	72.89
Males	82.96	80.89	91.90	90.00	93.78	74.67	79.05
Both	92.00	98.85	89.89	88.97	96.00	82.30	82.07

Table 4. The average percentage of identification accuracy for each emotion using AUSEES features and GMM.

Group	Emotions						
	Contempt	Angry	Anxious	Dysphoric	Pleasant	Neutral	Happy
Females	93.94	87.62	80.00	80.48	93.78	71.90	67.11
Males	87.41	87.11	90.00	75.24	88.44	48.44	80.00
Both	92.33	83.68	67.13	49.43	93.33	66.67	84.37

6. Conclusions

Two types of characteristic features (AUSEEG and AUSEES) based on the new nonlinear model of the glottal waveform formation were presented and tested in the process of automatic emotion classification in speech.

The results presented in Tables 2-4 showed significant gender based differences which confirmed previously reported similar observations [6,7]. Both classifiers showed similar trends, however the classification results obtained for the GMM were in most cases higher than for KNN.

Table 3 provides interesting gender based observations showing that for the glottal waveform (AUSEEG) all emotions except dysphoric and happy were easier to detect in female speakers than in male speakers. Table 4 on the other hand shows that with speech-based detection (AUSEES), all emotions except happiness were easier to detect in female speakers than in male speakers. As illustrated in the example of Fig. 3, dysphoric and happy affects exhibit very similar spectral distribution of energy which could explain the higher level of confusion between these two emotions.

Both types of the proposed features AUSEEG and AUSEES (Table 2) significantly outperformed the classical

MFCC features, indicating that the distribution of the spectral energy is an important factor in emotion discrimination.

It was also demonstrated (Table 2) that the new features AUSEEG representing the spectral energy distribution of the glottal waveform provided better classification rates than the AUSEES features representing the spectral energy distribution of the speech signal. This observation provides an important indication suggesting that most of the emotional aspect of speech is generated during the glottal wave formation, before the vocal tract filtering process.

7. Acknowledgements

This research was supported by the Australian Research Council Linkage Grant LP0776235. The authors would like to thank the ORYGEN Research Centre, Australia, the Oregon Research Institute, USA and Dr Lisa Sheeber for their invaluable help and support.

8. References

- [1] B. Davis, L. Sheeber, H. *et al.* "Adolescent Responses to Depressive Parental Behaviors in Problem-Solving Interactions", *J. of Abnormal Child Psychology*, 28(5) 2000.
- [2] G. Zhou, J.H.L. Hansen, J.F. Kaiser, 2001. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing*, IEEE Transactions on 9(3): 201-216.
- [3] S. Khosla, S. Murugappan, R. Paniello, J. Ying, E. Gutmark, "Role of vortices in voice production: Norma versus asymmetric tension", *The Laryngoscope*, 119, January 2009, pp. 216-221.
- [4] H. Teager, Some observations on oral air flow during phonation. *Acoustics, Speech and Signal Processing*, IEEE Transactions on, 1980. 28(5): p. 599-601.
- [5] A. Barney, *et al.* "Fluid flow in a dynamic mechanical model of the vocal folds and tract", *JASA*.1999, 105(1), pp. 444-455.
- [6] E. Moore, M. A. Clements, *et al.* "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech." *IEEE T. Bio Eng*, 2008, 55(1): 96-107.
- [7] L.S.H. Low, M. Lech, *et al.* Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents. in *Cognitive Informatics*, 2009. ICCI '09..
- [8] D.K. Ververidis, C.Kotropoulos, Emotional speech recognition: Resources, features, and methods. *Speech Communication* (48) (2006) 1162–1181.
- [9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion Recognition in human-computer interaction, *IEEE Signal Processing Magazine* 18(1) (2001) 32-80.
- [10] S. Khosla, S. Murugappan, E. Gutmark, What can vortices tell us about vocal vibration and voice production, *Current opinion in Otolaryngology & Head and Neck Surgery*, (16) (2008).
- [11] W. Zhao, C. Zhang, SH Frankel, L. Mongeau, Computational aeroacoustics of phonation, part 1: computational methods and sound generation mechanisms. *J Acoust Soc Am* 2002;112:2134–2146.
- [12] D. Shinwari, RC Scherer, A. Afje, K. Dewitt, Flow visualization in a model of the glottis with a symmetric and oblique angle. *JASA* 2003;113:487–497.
- [13] T. Quatieri, *Speech Signal Processing*, Prentice Hall 2002.
- [14] Veeneman, D. and S. BeMent, Automatic glottal inverse filtering from speech and electroglottographic signals. *Acoustics, Speech and Signal Processing*, IEEE Transactions on, 1985. 33(2): p. 369-377.
- [15] He L., *et al.* "Emotion Recognition in Spontaneous Speech within Work and Family Environments", *iCBBE* 2009, Jun 11th to 13th, 2009 in Beijing, China.