Exploring the Significance of Low Frequency Regions in Electroglottographic Signals for Emotion Recognition

S.G. Ajay(✉), D. Pravena, D. Govind, and D. Pradeep

Centre for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Coimbatore 641112, Tamilnadu, India
ajay190694@gmail.com, d.pravena@gmail.com, d_govind@cb.amrita.edu,
getpradeepd@gmail.com
http://www.amrita.edu/campus/coimbatore

Abstract. Electroglottographic (EGG) signals are acquired directly from the glottis. Hence EGG signals effectively represent the excitation source part of the human speech production system. Compared to speech signals, EGG signals are smooth and carry perceptually relevant emotional information. The work presented in this paper includes a sequence of experiments conducted on the emotion recognition system developed by the Gaussian Mixture Modeling (GMM) of perceptually motivated Mel Frequency Cepstral Coefficients (MFCC) features extracted from the EGG. The conclusions drawn from these experiments are two folds. (1) The 13 static MFCC features showed improved emotion recognition performance than 39 MFCC features with dynamic coefficients (by adding Δ and $\Delta\ \Delta$). (2) Low frequency regions in the EGG are emphasized by increasing the number of Mel filters for MFCC computation found to improve the performance of emotion recognition for EGG. These experimental results are verified on the EGG data available in the classic German emotional speech database (EmoDb) for four emotions such as (Anger, Happy, Boredom and Fear) apart from Neutral signals.

Keywords: EGG · MFCC · GMM · HTK · openEAR

1 Introduction

With the advancements in Machine Learning and Artificial Intelligence, the need for effective Human-Machine interaction has gained significant importance. The impact of emotional speech in Human-Machine interaction is less significant due to the fact that, machines cannot understand human's emotional state [13]. This has increased the need for analysis of emotions during the Human-Machine interaction. Usually, emotions of human beings are analyzed from the recorded speech signals. Rather than using recorded speech signals, Electroglottographic

S.M. Thampi et al. (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 678,
DOI 10.1007/978-3-319-67934-1_28

(EGG) signals can be used for recognizing the emotions. Other than EGG signals, another approximation representing the glottal information (excitation source) is by using linear prediction residual signals [2,5], which can be derived from speech signals. By the informal listening to EGG signals, humans can identify the emotion it carries. With the availability of EGG data in different emotions in the classic German emotional speech database (EmoDb), the present work focuses on using EGG signals for emotion recognition.

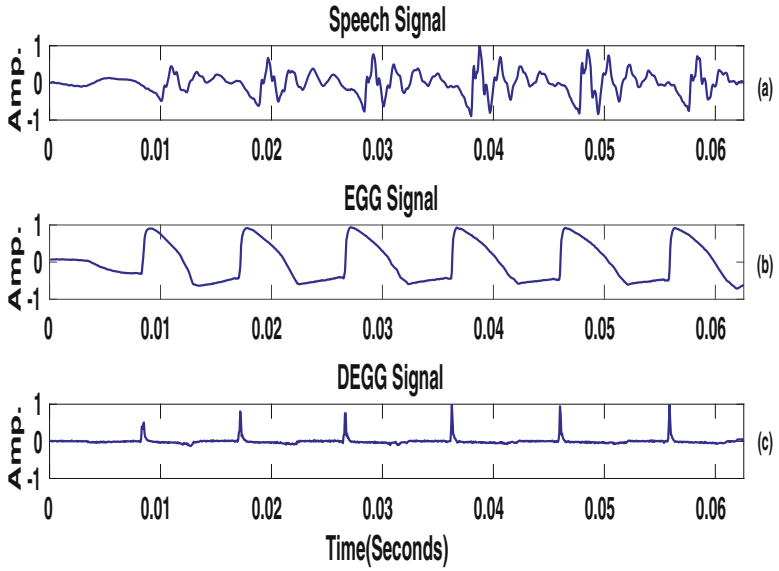

Fig. 1. Smooth nature of EGG signal. Plot (a) Speech signal, (b) EGG signal, (c) Differenced EGG signal.

Figure 1 clearly depicts the difference between the speech and the EGG signals. As seen from the Fig. 1, EGG signals are smooth compared to speech signals and also when they are differentiated, a sequence of impulses are produced which represents the glottal closure instants (GCIs). This indicates that the glottal information is clearly present in it and also it carries perceptually relevant emotional information (excitation source information) in the low frequency regions. This phenomenon is shown clearly in the Fig. 2. The analysis of this phenomenon is performed by calculating wide-band spectrogram for different emotional (Anger, Happy, Boredom, and Fear) utterances apart from Neutral signals present in the EGG data. The same vowel is chosen for representing the variations in different emotions through the spectrogram. Figure 2(f)–(j) shows the spectrogram of the EGG signal produced when the vowel /a/ is elicitated with the different emotional state of the same speaker. The corresponding EGG signal is plotted in Fig. 2(a)–(e). In all the spectrograms, the low frequency regions are very dark when compared to the high frequency regions, which shows that the

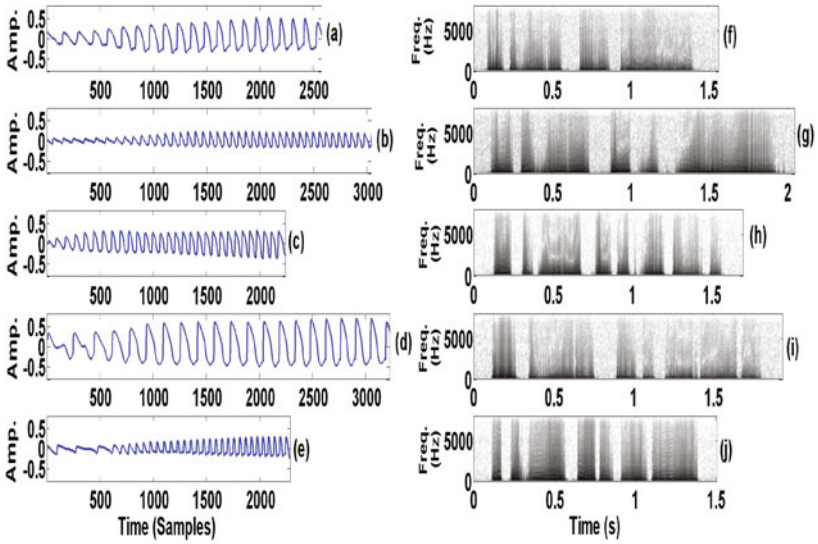

Fig. 2. Spectrogram analysis of glottal signals for the same utterance with different emotions. Plot (a)–(e) shows the glottal signals for Neutral, Anger, Happy, Boredom and Fear emotions respectively. Plot (f)–(j) are the corresponding wide-band spectrogram of glottal signals in (a)–(e).

availability of emotional information in glottal signals are more concentrated in the low frequency regions.

In this work, the state of the art perceptually motivated Mel Frequency Cepstral Coefficients (MFCC) are considered as features of the EGG signals for Gaussian Mixture Modelling [10]. Since the human perception of sound is in Mel scale, Mel filters were used for the computation of MFCC features in speech signals. MFCC's are widely used as features for many applications like Speaker Recognition, Speaker Identification, Speech Recognition, Emotion Recognition etc. [9,14,15]. The works presented so far in the literature shows that, emotion recognition is performed exclusively for emotive speech signals [1,7,11,16]. Pati et al. [12] used Residual MFCC (RMFCC) features from the linear prediction residual signals which is an approximation of glottal signals derived from the speech. Since the EGG signals have only glottal information, the proposed emotion recognition system uses MFCC features from it. So the work presented in this paper, attempts to experiment on the EGG signals by extracting the 13 static MFCC and the 39 dynamic MFCC features, by varying the number of filters in the Mel filter bank in order to emphasize the low frequency regions for computing MFCC. The organization of the work is as follows: Sect. 2 refers to the Development of emotion recognition system using EGG. Section 3 explains the Performance analysis of Emotion recognition using EGG signals. Summary and Conclusion of the present work are discussed in Sect. 4.

2 Development of Emotion Recognition System Using EGG

2.1 Production of EGG

EGG encompasses more emotional information which is captured at the time of elicitation. It is recorded through a device named Electroglottograph as shown in Fig. 3, which contains a pair of electrodes placed near the glottis region to capture the vocal fold vibrations during the production of speech [8]. It measures the vibration by passing a small amount of current between the contact area of the vocal folds. The impedance across the electrodes varies with respect to the vibrations of the vocal folds [6]. This variation of impedance produces quasi-periodic (non-stationary) signals known as the EGG signals.

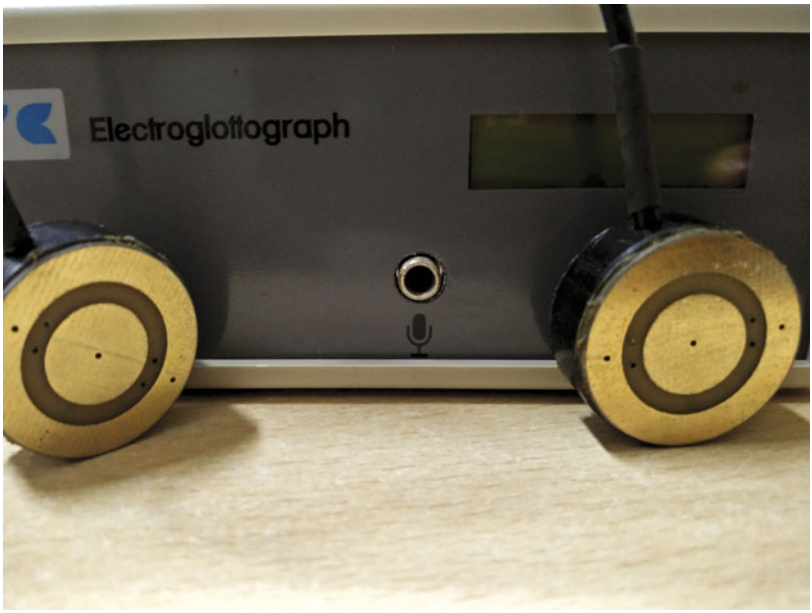

Fig. 3. Electroglottograph.

MFCC Feature Extraction from EGG. A nonlinear triangular Mel scale filter bank [14] as shown in Fig. 4 (filters are linearly placed in the low frequency regions (<1000 Hz) and logarithmically placed in the high frequency regions (>1000 Hz)) has the potentiality to emphasize the lower frequency components over the higher ones. Mel filters are designed to mimic human auditory perception of sound by concentrating more on the low frequency regions. As EGG signals are low frequency in nature, Mel Frequency Cepstral Coefficients can act

as good features representing the emotional information present in the low frequency regions. In order to extract features, the non-stationary signal is divided into a smaller number of stationary frames of size 20 ms using Hamming window. Hamming window is used to avoid spectral leakage. A Hamming window shift of 10 ms is used. Along with the 13 static MFCC (velocity features), Δ and $\Delta\ \Delta$ (acceleration features) are extracted from each frame and they are combined with the 13 MFCC features to make the feature vector dimension as 39. By increasing the filter banks ranging from (14, 16, 18....46), 13 and 39 MFCC features are extracted individually.

3 Performance Analysis of Emotion Recognition Using EGG Signals

The performance of emotion recognition using EGG signals is analyzed in the classic German emotional speech database (EmoDb) [3] which includes a simultaneous recording of Speech and EGG signals. The database was developed for six different emotions Anger, Happy, Fear, Boredom, Sad, Disgust apart from Neutral signals with 10 professional actors (5 Male and 5 Female) using 10 neutral sentences spoken in six emotions. Out of six emotions, four emotions (Anger, Happy, Fear, Boredom) along with Neutral signals are considered for this emotion recognition analysis. Each speech sample is recorded at a sampling rate of 48 KHz with 16 bits per sample resolution. In this work Speech and EGG signals of German emotional speech data are separated, and the separated EGG signals are downsampled to 16 KHz. Training and Testing for the analysis are performed with 590 utterances. Out of 590 utterances, 474 utterances were taken for training the GMM's and 116 utterances were used for testing the GMM's. A series of experiments were conducted with the 13 static MFCC features and the 39 dynamic MFCC features with different filter bank coefficients. These cepstral features are trained with 512 Gaussian Mixture components as the training data is small and the trained GMM's are tested for the classification accuracy in different emotional classes. For implementing MFCC-GMM based emotion recognition system, we have used HTK (Hidden Markov Model) toolkit [17] in our experiments. The configuration files are given in the following link https://drive.google.com/drive/folders/0BzHkgLdbz2n-OGR5dnJHTXhwVTg?usp=sharing. From the experiments conducted the observations inferred are two folds.

EGG signals show better performance in classifying the emotions with the 13 static MFCC features than the 39 dynamic MFCC features (by adding Δ and $\Delta\ \Delta$) for the conventional Mel filter bank of size 28 as seen from Table 1.

The rationality in this performance is due to the fact that, while taking the 39 dynamic MFCC features the change in the dynamics of the vocal tract across different frames of an audio signal is accounted. Unlike speech signals, EGG signals contain only glottal information (excitation source information) and it is clearly captured by the 13 static MFCC features. Since EGG signals lacks the vocal tract information, accounting the dynamic features (Δ and $\Delta\ \Delta$) which

Table 1. Classification Accuracies(%) for emotion recognition in EGG from German EmoDb with the 13 and 39 MFCC features for the conventional filter bank of size 28.

Number of Gaussians	Accuracy(%)	
	Filter bank of size 28	
	13 MFCC	39 MFCC
8	49.14	54.31
16	64.66	56.03
32	60.34	58.62
64	71.55	63.79
128	76.72	67.24
256	75.86	67.24
512	77.59	68.10

represents the same is not helping in improving the recognition performance. Therefore the proposed work experiments only on the use of the 13 static MFCC features by increasing the number of filters in the low frequency regions, thereby giving more emphasis.

Tables 2 and 3 discusses the series of experiments conducted with the 13 static MFCC features by varying the number of Mel filters in the Mel filter bank.

Table 2. Classification Accuracies(%) for emotion recognition in EGG from German EmoDb with the 13 static MFCC features containing different filter bank coefficients.

Number of Gaussians	Accuracy(%)										
	Different sizes of Mel filter banks										
	14	**16**	**18**	20	22	24	26	28	30	32	34
8	54.31	52.59	50.00	50.00	50.86	52.59	49.14	49.14	51.72	51.72	52.59
16	59.48	56.03	62.93	59.48	62.93	58.62	63.79	64.66	61.21	64.66	62.93
32	65.52	68.10	67.24	67.24	65.52	66.38	68.10	60.34	65.52	65.52	66.38
64	70.69	67.24	70.69	68.97	72.41	71.55	70.69	71.55	71.55	70.69	69.83
128	74.14	72.41	77.59	73.28	71.55	75.86	71.55	76.72	75.86	75.86	74.14
256	**72.41**	**72.41**	**74.14**	76.72	72.41	77.59	72.41	75.86	78.45	76.72	72.41
512	69.83	73.28	75.00	76.72	73.28	76.72	76.72	77.59	79.31	77.59	75.86

It is evident from Tables 2 and 3 that, performance of the MFCC-GMM based emotion recognition system increases by increasing the number of Mel filters in the low frequency regions. This is because, EGG signals are low frequency in nature and therefore by keeping more filters in the low frequency regions, more emotional information is captured. The optimal performance with 80.17% is obtained while using 256 Gaussian mixtures with the 13 static MFCC features for the higher order filter bank of size 38. Also when the filters are increased

Table 3. Classification Accuracies(%) for emotion recognition in EGG from German EmoDb with the 13 static MFCC features containing different filter bank coefficients.

Number of Gaussians	Accuracy(%)					
	Different sizes of Mel filter banks					
	36	**38**	**40**	**42**	44	46
8	51.72	50.86	49.14	51.72	46.55	43.97
16	63.79	60.34	62.93	62.07	62.93	61.21
32	61.21	61.21	66.38	66.38	59.48	60.34
64	69.83	73.28	70.69	71.55	73.28	72.41
128	73.28	76.72	75.86	75.86	77.59	75.00
256	76.72	**80.17**	**79.31**	**79.31**	75.86	75.00
512	75.00	77.59	77.59	77.59	76.72	79.31

beyond 38, the recognition performance seem to degrade, this is due to the fact that, when filters are more denser at the low frequency regions, the width of the traingular filters decreases, this, in turn, fails to capture the relevant emotional information. Figure 4 shows the traingular Mel filter bank of size 28 and 36. It is evident from the Fig. 4 that, when filters are denser in the lower frequency regions more emphasis is given.

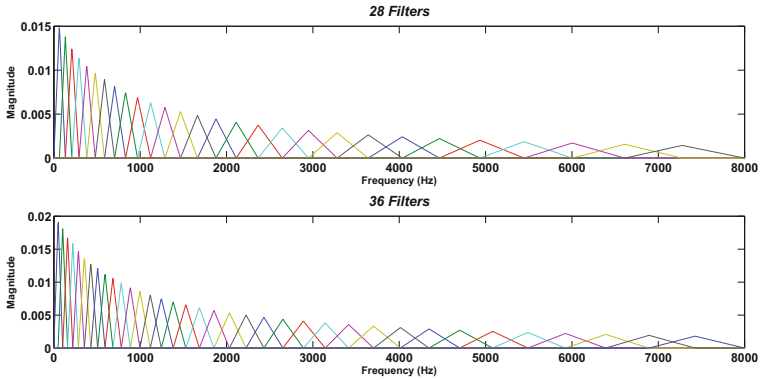

Fig. 4. Mel Filter banks of size 28 and 36.

4 Summary and Conclusion

The work proposed in this paper focuses on using the EGG signals for emotion recognition. As EGG signals are low frequency in nature and it approximates the glottal information during the production of the emotive speech signals,

the perceptually motivated Mel Frequency Cepstral Coefficients (MFCC) are extracted from the same for Gaussian Mixture Modeling. The conclusions drawn from this work is as follows,

- The MFCC-GMM system with the 39 dynamic MFCC features with Δ and $\Delta\ \Delta$ does not contribute for improved emotion recognition performance in case of the EGG signals, whereas the 13 static MFCC features give better performance for the same.
- In order to emphasize the low frequency components, increasing the number of Mel filters in Mel filter bank to a certain level for the computation of MFCC features helps to improve the emotion recognition performance.

The future work concentrates on building an emotion recognition system using the acoustic features derived from the Munich's openEAR toolkit [4] for the same EGG signals present in the classic German emotional speech database (EmoDb). The results obtained for the emotion recognition from the conventional state of the art MFCC-GMM can be verified or improved using other classification algorithms in Deep Networks.

References

1. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Spoken emotion recognition using hierarchical classifiers. Comput. Speech Lang. **25**, 556–570 (2011)
2. Ananthapadmanabha, T.V., Yegnanarayana, B.: Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Sig. Process. **27**(4), 309–319 (1979)
3. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlemeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of INTERSPEECH, pp. 1517–1520 (2005)
4. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor, pp. 1459–1462 (2010)
5. Govind, D., Prasanna, S.R.M.: Expressive speech synthesis: a review. Int. J. Speech Technol. **16**(2), 237–260 (2013)
6. Henrich, N., DAlessandro, C., Doval, B., Castellengo, M.: On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. J. Acoust. Soc. Am. **115**(3), 1321–32 (2004)
7. Kandali, A.B., Routray, A., Basu, T.K.: Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In: IEEE Region 10 Conference (2008)
8. Kitzing, P.: Clinical applications of electroglottography. J. Voice **4**(3), 238–249 (1990)
9. Koolagudi, S.G., Rao, K.S.: Two stage emotion recognition based on speaking rate. Int. J. Speech Technol. **14**, 35–48 (2011)
10. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech using source, system, and prosodic features. Int. J. Speech Technol. **15**, 265–289 (2012)
11. Neiberg, D., Elenius, K., Laskowski, K.: Emotion recognition in spontaneous speech using GMMS. In: INTERSPEECH (2006)

12. Pati, D., Prasanna, S.R.M.: Processing of linear prediction residual in spectral and cepstral domains for speaker information. Int. J. Speech Technol. **18**(3), 333–350 (2015)
13. Prasanna, S.R.M., Govind, D.: Analysis of excitation source information in emotional speech. In: Proceedings INTERSPEECH, pp. 781–784 (2010)
14. Pravena, D., Nandhakumar, S., Govind, D.: Significance of natural elicitation in developing simulated full blown speech emotion databases, pp. 261–265 (2016)
15. Raviram, P., Umarani, S.D., Wahidabanu, R.S.D.: Isolated word recognition using enhanced MFCC and IIFS. In: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), vol. 199, pp. 273–283. Springer (2013)
16. Vondra, M., Vch, R.: Recognition of emotions in German speech using Gaussian mixture models. Multimodal Sig. **5398**, 256–263 (2009)
17. Young, S.J., Young, S.: The HTK hidden Markov model toolkit: design and philosophy (1993)