Analysis of Emotional Speech using Excitation Source Information

P. Gangamohan

Speech Processing Laboratory, KCIS
International Institute of Information Technology, Hyderabad, India
gangamohan.p@students.iiit.ac.in

1. Evolution of the thesis

The human speech production mechanism produces acoustic signals corresponding to one of the three broad categories: Neutral speech, emotional speech and non-verbal sounds.

- Neutral speech is produced in non-emphatic state of the speaker. It is generally inferred to carry a sequence of linguistically meaningful sound units.
- Emotional speech consists of superposition of emotion characteristics on (neutral) speech, in such a way that both the linguistic message and emotion characteristics are conveyed through the signal.
- Non-verbal sounds do not carry linguistic information, but are produced by the speech production mechanism. These sounds are produced either by voluntarily controlling the production process as in artistic voices, or by involuntary/spontaneous production as in the case of sounds, such as laughter, scream and cry.

Speech systems, such as speech recognition and speaker recognition, are developed using a parametric representation of neutral speech. The performance of these systems may degrade when they are tested with the parametric representation of emotional speech. This indicates that there are certain changes in the speech production while producing sounds in emotional state, and these changes reflect in the parametric representation. Moreover the same parametric representation is used for discriminating emotional speech from neutral speech, and discriminating among emotion categories. Whereas it is necessary to determine the production features that contribute to emotion, and extract those features from the speech signal for discriminating different emotion categories (i.e., emotion recognition).

In approaches for emotion recognition, a classification model is developed on the features extracted. Typically, these features are broadly categorized into voice quality, prosody and spectral features. The subsegmental and suprasegmental information related to the excitation source component is given in voice quality and prosody features, respectively [1, 2, 3, 4]. The segmental information related to the vocal tract system component is given in spectral features, such as mel-cepstral coefficients, linear prediction cepstral coefficients and perceptual linear prediction coefficients [5, 6]. Although these features explain the production of emotional speech to some extent, they carry significant information of speaker's signature and sound unit, and they share similar properties among emotions. Thus requiring a phonemically balanced database of several speakers for developing a classification model to cover speaker and sound unit variability. It is a very difficult task to build such a database. Due to continuum nature of most of the emotions, it is difficult to obtain ground truth for developing emotion recognition systems [7, 8].

The primary objective of this thesis is to understand the production characteristics responsible for emotional speech, and to identify and extract emotion-specific features to discriminate different emotion categories. Since emotion characteristics are embedded with the linguistic message and speaker's signature, it is a challenge to isolate emotion-specific features from linguistic message and speaker characteristics in the speech signal.

The first study carried out in this thesis is to examine the significance of different components of speech production that contribute to the perception of emotion in speech. For this purpose, a flexible analysis-synthesis tool (FAST) is developed to selectively control the different components [9]. It is observed that the emotion characteristics are reflected more in the excitation source components, and to a lesser extent in the segmental information reflecting the shape of the vocal tract [10]. Thus the focus of the following studies is to explore the excitation-related features to represent the emotion characteristics present in the speech signal.

During speech production, the significant excitation of the vocal tract system takes place due to vocal fold vibration. Within each glottal cycle, the impulse-like excitation at the glottal closure instant (GCI) gives high signal-to-noise ratio (SNR) characteristics in the speech signal. Features extracted around the GCI locations may carry emotion related information. In order to extract these subsegmental features, identification of GCIs from the speech signal plays a vital role.

Most of the existing GCI extraction methods give better results on neutral speech, but their performance degrade in the case of emotional speech. In this thesis, a robust and alternative approach to zero frequency filtering (ZFF) method is proposed in order to overcome some issues. The ZFF method exploits the properties of impulse-like excitation by passing a speech signal through the resonator whose pole pair is located at 0 Hz, on the unit circle, in the z-plane. As the resonator is unstable, the polynomial growth/decay is observed in the filtered signal, thus requiring a trend removal operation. It is observed that the length of the window for trend removal operation is critical in speech signals where there are more fluctuations in the fundamental frequency (F_0). The proposed finite impulse response (FIR) filter is designed by placing a large number of zeros at $\frac{fs}{2}$ Hz, closer to the unit circle [11]. This is a stable filter where the output signal does not have polynomial growth/decay as in the case of ZFF method. Furthermore a priori knowledge of the local pitch period is not that critical in this method.

Some excitation based features extracted around the GCIs are examined for their significance in discriminating emotions [12]. The excitation source features considered are: Instantaneous F_0, strength of excitation (SoE), energy of excitation (EoE) and loudness parameter (η). These features are extracted from the Hilbert envelope of the linear prediction (LP) residual signal and ZFF signal. A sample speech waveform, its Hilbert envelope of the LP residual signal and ZFF signal are given in Fig. 1. The 2-D feature distributions (scatter plots) for neutral and angry utterances are shown in Fig. 2. A template matching

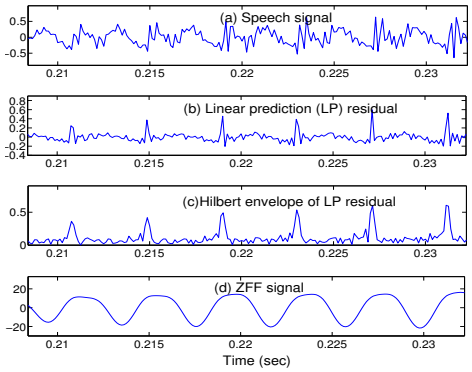

Figure 1: *(a) Speech signal, (b) Linear prediction (LP) residual, (c) Hilbert envelop of LP residual, and (d) Zero frequency filtered (ZFF) signal*

based emotion classification is developed using the normalized feature distributions. As the dynamic range of these features are speaker dependent, the distributions are normalized with respect to neutral speech. Thus requiring a reference neutral speech for using these features.

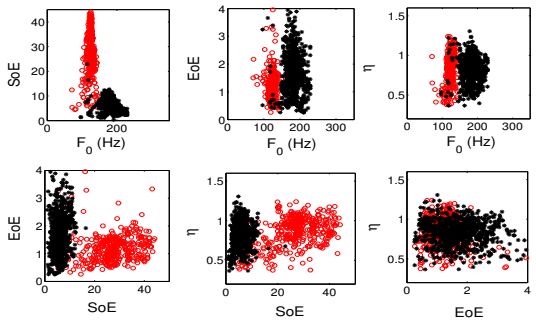

Figure 2: *The 2-D feature spaces for a speaker's neutral utterance (marked by 'o') and angry utterance (marked by '*').*

Humans can discriminate emotions without having a reference (in most of the cases). For exploring features that do not depend on a reference, other voice categories which describe the perception of emotion characteristics, such as arousal and valence, are considered. It is also realized that the feature sets might be different for different categories, and hence it is necessary to consider a hierarchical approach for emotion studies as shown in Fig. 3. Speech produced by a speaker without any physical and mental disorders is dealt in this thesis. The following are the studies identified and being carried out:

- Discrimination between modal speech and falsetto speech.
- In modal speech, discrimination between high arousal (angry, happy, afraid) speech and low arousal (neutral, sad, boredom) speech.
- In high arousal modal speech, discrimination between angry speech and happy speech.

The assumption is that the excitation source information is much different in all the above voice categories.

Firstly, the glottal vibration characteristics around the GCIs using electroglottograph (EGG) signals are examined. Within a glottal cycle, the magnitude of the differenced EGG (dEGG) signal at the GCI has a sharp gradient. This is due to the abruptness in closing mechanism of vocal folds. When compared to modal speech, falsetto speech has more spread out in the samples of the dEGG signal around the GCIs. This indicates that there is smoother closing of vocal folds at the GCI, as shown in Fig. 4 (where the segments of the dEGG signal around the GCIs are superimposed). Extraction of this subsegmental information around the GCIs is substantial to discriminate falsetto and modal categories. But it is a challenge to extract this information from the speech signal.

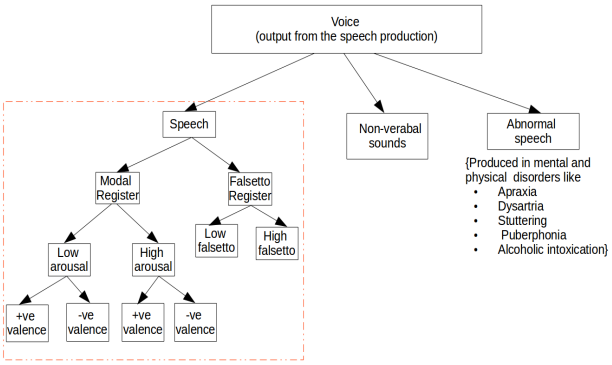

Figure 3: *A hierarchical approach for emotion studies.*

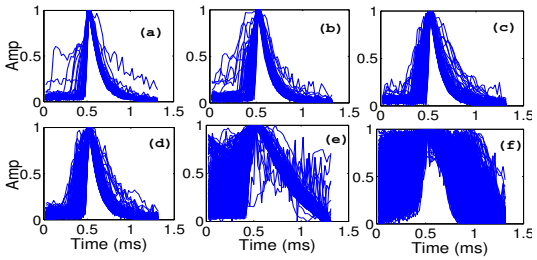

Figure 4: *Illustration of glottal vibration characteristics using the EGG signals. (a) neutral, (b) happy, (c) angry, (d) shout, (e) low-falsetto, and (f) high-falsetto utterances.*

The excitation source information in an entire glottal cycle might give significant information of high arousal speech. Among discrimination of high arousal categories like anger and happiness, it appears that the suprasegmental excitation based features play a key role.

All these studies demonstrate two aspects: There is a need to consider different feature sets for different tasks in emotion studies. The significance of analysis of vocal fold vibration characteristics, which is the primary mode of excitation, and also which contributes to the production of different types of emotions.

2. Acknowledgements

I express sincere gratitude to my advisors Prof. B. Yegnanaryana and Dr. Suryakanth V Gangashetty for guiding me throughout my PhD. Also, I would like to thank Tata Consultancy Services (TCS) for partially funding my PhD programme.

3. References

[1] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[2] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[3] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[4] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Trans. Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011.

[5] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[6] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," in *Speech Evaluation in Psychiatry*, J. K. Darby, Ed. Grune and Stratton, New York, 1981.

[7] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. Int. Conf. Multimedia and Expo*, Hannover, Germany, 2008, pp. 865–868.

[8] Khiet P. Truong, David A. van Leeuwen, and Franciska M. G. de Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech Communication*, vol. 54, no. 9, pp. 1049–1063, 2012.

[9] P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "A flexible analysis and synthesis tool (FAST) for studying the characteristic features of emotion in speech," in *Proc. IEEE Int. Conf. Consumer Communications and Networking Conference*, Las Vegas, USA, 2012, pp. 266–270.

[10] P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "Relative importance of different components of speech contributing to perception of emotion," in *Proc. Speech Prosody*, Shangai, China, 2012, pp. 657–660.

[11] P. Gangamohan and B. Yegnanarayana, "A robust and alternative approach to zero frequency filtering method for epoch extraction," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017.

[12] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 1916–1920.