EXTRACTION OF EXCITATION INFORMATION FROM SPEECH AND ITS APPLICATIONS FOR EXPRESSIVE SPEECH PROCESSING

Sudarsana Reddy Kadiri[1] *and* **Advisor:** *B. Yegnanarayana*[2]

[1]International Institute of Information Technology, Hyderabad-500032, India.
[2]Birla Institute of Technology & Science Pilani, Hyderabad-500078, India.
sudarsanareddy.kadiri@research.iiit.ac.in;yegna@iiit.ac.in

ABSTRACT

Through speech production mechanism, speech with different voice qualities such as phonations, emotions, expressive singing and other paralinguistic sounds are also produced. Most of these sounds demonstrate these features mostly due to the excitation component (vibration of the vocal folds at the glottis) whereas the dynamic vocal tract system primarily conveys the message. Hence, the excitation source processing acquires significance especially for the analysis, detection and representation of expressive voices. Most of the existing excitation source information extraction methods are not reliable especially when applied on expressive voices, mainly due to significant source-system coupling. Hence, there is a need for new signal processing methods that can capture the dynamic variations in excitation source so that different types of sounds can be better analyzed and represented. The objective of this work is to derive new signal processing methods to extract the excitation source information directly from the signal and then investigate the significance of this information for the analysis and detection of various expressive voices. Towards this, some of the excitation source features are extracted using recently proposed signal processing methods and then the significance of these excitation features are studied in emotional speech analysis and recognition. Presently the studies are in progress in representing the excitation source in the form of impulse-like sequence.

Index Terms— Signal processing, speech analysis, excitation source, expressive speech processing, emotion recognition.

1. INTRODUCTION

Speech signal contains rich information about message (lexical content), language, dialect, gender, age, speaker characteristics, personality and emotion/expressive state. In this work, our focus is on analyzing and recognizing the expression/emotion present in the speech signal. It is known that, the dynamic variations in the excitation source component significantly contribute to the production and perception of various voice qualities, emotions/expressions in speech. Hence, extraction of glottal source information gained significance, especially for deriving the glottal source waveform and characterizing it for various applications such as expressive speech analysis, speech synthesis, speaker recognition, etc [1, 2].

Most of the existing methods for excitation source information extraction depend on the inverse filtering of speech for estimating the glottal source waveform. Features derived from glottal source waveform such as normalized amplitude quotient (NAQ), open quotient (OQ), closing quotient (CQ), difference between first and second harmonics (H1-H2) etc., in vowel segments are analyzed in phonation types and emotional speech ([3], and the references therein). From the analysis, it was observed that NAQ and H1-H2 are shown to provide good discrimination. In the case of emotional speech, NAQ is shown to provide better correlation with arousal (active or passive) rather than valence (positive or negative) for both genders. Even though NAQ correlates with emotions, it is to be noted that NAQ by itself is not sufficient to discriminate the emotions accurately. Some studies also investigated the interdependencies among the excitation source features in sustained vowels of emotional speech [4]. It is known from the literature that, the current inverse filtering methods are not sufficient for the estimation of glottal source waveform from continuous speech and the effect on these methods is severe especially in the expressive voices due to rapid high pitch variations and significant source-filter coupling [3].

There are some attempts in the literature to extract some specific excitation features from speech signal. The important features among them are glottal closure instant (GCI), glottal opening instant (GOI), strength of glottal closure, sharpness of glottal closure, fundamental frequency and various phases in a glottal cycle [2, 5]. For extracting most of these features, existing methods use linear prediction (LP) residual (approximate excitation signal) ([2], and the references therein). It was noticed that extraction of these features from expressive voices are not reliable as extracting the excitation signal for such voices is difficult. To address these issues, some attempts were made in the literature such as extraction of pitch in highly expressive voice like noh singing voice [6]. In this work, we are focusing on extracting features directly from speech signal without depending on LP residual/excitation signal.

2. OBJECTIVES OF THE WORK

The primary objective of the work is to extract the excitation source information from the continuous speech directly without depending on characteristics of vocal tract system. The secondary objective of the work is to investigate the significance of the excitation source information for the analysis and detection of various expressive voices.

3. HYPOTHESIS OF THE WORK

The primary mode of excitation during production of speech is due to the vibration of the vocal folds at the glottis. Even though the excitation information is present throughout the glottal cycle, it is considered to be significant only when there is a large change in short-time interval, i.e., when it is impulse-like. Therefore, the glottal excitation can be approximated as a sequence of impulses (hypothesis). This approximation on the excitation of the vocal-tract system suggests a new approach for processing the speech signal. One feature of the impulse-sequence of excitation is the GCI and it is the most significant excitation which takes place at the instant of glottal closure (as discussed in Sec. 4.1).

4. RESEARCH CARRIED OUT

In this section, first we discuss the extraction of major impulse-like excitation/glottal closure instant in expressive voices such as emo-

tional speech [7] and singing voice [8]. Then, the significance of the excitation source features is discussed for the analysis and recognition of emotional speech [9, 10].

4.1. Extraction of major impulse-like excitation/glottal closure instant (GCI) from singing voice

Existing methods of GCI detection are not suitable for handling voices which have rapid variations in pitch and significant source-filter coupling such as singing/emotional speech. To overcome this, a modified zero frequency filtering method for epoch extraction was proposed in [8]. In this, the differenced signal in short segments of around 0.4 to 0.5 sec is passed through a cascade of three zero frequency resonators (ZFR). The trend in the output of ZFR is removed by subtracting the local mean computed over the average pitch period at each sample. The resulting signal is called as modified zero frequency filtered (modified ZFF) signal and the instants of negative-to-positive zero crossings of modified ZFF signal correspond to the GCIs [8, 11]. As illustrated in Fig. 1, there is a close agreement between negative peaks of the differenced EGG (dEGG) signal and the GCIs obtained from the modified ZFF signal.

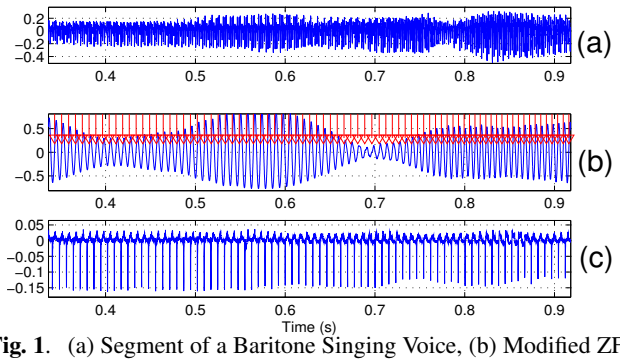

Fig. 1. (a) Segment of a Baritone Singing Voice, (b) Modified ZFF signal (epoch locations marked by arrows), and (c) dEGG signal [8].

4.2. Emotional speech analysis and recognition

The significance of the excitation source features (such as fundamental frequency (F_0), strength of the impulse-like excitation (SoE) and energy of excitation (EoE)), derived from continuous speech signal were studied in the case of emotional speech analysis and recognition [9]. These features are derived around epochs using ZFF and LP analysis. The deviations in the features of emotional speech w.r.t neutral speech were analyzed in three 2-dimensional distributions, as shown in Fig. 2. In this, it can be observed that there are significant deviations in emotional features compared to neutral speech.

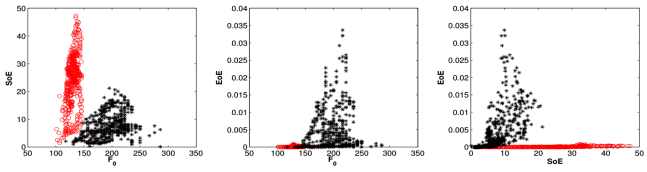

Fig. 2. *Three combinations of 2-D feature pairs for a male speaker neutral utterance ('o') and emotion (happy) utterance ('*') [9].*

In order to characterize these deviations Kullback-Leibler (KL) distance is computed between the corresponding 2-D feature distributions. Based on the deviations of emotional speech features with reference to neutral speech features, an emotion recognition system is developed [9]. The results of emotion recognition using the excitation features are shown in Table 1.

From the results, it can be observed that the excitation source features capture emotion specific information. It can also be observed that, there is a scope in improving the performance especially by increasing the discrimination of anger and happy emotions [12].

Table 1. *Emotion recognition results achieved on EMO-DB using excitation source features [9].*

	Neutral	Sad	Anger	Happy
Neutral	74/79	4/79	0/79	1/79
Sad	27/62	33/62	0/62	2/62
Anger	2/127	0/127	114/127	11/127
Happy	7/71	3/71	27/71	34/71

5. WORK IN PROGRESS

Currently we are working on representation of speech using impulse-like sequence. For this, we are exploring a recently proposed signal processing method, single frequency filtering (SFF) [13]. In the SFF approach, the instantaneous amplitude and phase components of the speech signal are obtained at any desired frequency by frequency shifting the signal and filtering the resulting signal using an all-pole filter. The root of the all-pole filter is located on the unit circle at the highest frequency, i.e., at $f_s/2$, where f_s is the sampling frequency. It was observed that SFF provides high resolution of spectral features such as sharper harmonics and high resolution of temporal features like impulses. Features derived from SFF method were shown to be useful for speech/nonspeech discrimination [13], fundamental frequency estimation [14], GCI detection [7], time delay estimation [15], improving intelligibility [16], etc. The significance of the phase component of SFF for reconstruction of speech signals is also investigated [17].

6. ACHIEVED AND EXPECTED CONTRIBUTIONS

We proposed GCI detection method for emotional speech and singing voice which can handle rapid variations in pitch and significance of source-system coupling [7, 8]. Also, we derived excitation source features and analyzed the significance of these features in emotional speech analysis [10]. Further, an approach for emotion recognition is developed by capturing the deviations of emotion features from neutral speech features [9, 12]. Presently, we are working towards representing the excitation source in terms of impulse-like sequence (impulses even within a glottal cycle). This representation suggests a new approach for processing the speech signals capturing the dynamic variations in the excitation source, which may be useful for analyzing various expressive voices also.

References

[1] Thomas F Quatieri. *Discrete-Time Speech Signal Processing*. Pearson education, Singapore, 2004.

[2] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech and Language*, 28(5):1117–1138, 2014.

[3] Paavo Alku. Glottal inverse filtering analysis of human voice production-A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011.

[4] Johan Sundberg, Sona Patel, Eva Björkner, and Klaus R. Scherer. Interdependencies among voice source parameters in emotional speech. *IEEE Trans. on Affective Computing*, 2(3):162–174, 2011.

[5] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. Gerratt, J. Neubauer, and A. Alwan. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *Journal of the Acoustical Society of America*, 132:2625–2632, 2012.

[6] Osamu Fujimura, Kiyoshi Honda, Hideki Kawahara, Yasuyuki Konparu, Masanori Morise, and J. C. Williams. Noh voice quality. *Logopedics Phoniatrics Vocology*, 34(4):157–170, 2009.

[7] Sudarsana Reddy Kadiri and B. Yegnanarayana. Epoch extraction from emotional speech using single frequency filtering approach. *Speech Communication*, 86:52 – 63, 2017.

[8] Sudarsana Reddy Kadiri and B. Yegnanarayana. Analysis of singing voice for epoch extraction using zero frequency filtering method. In *ICASSP*, pages 4260–4264, April 2015.

[9] Sudarsana Reddy Kadiri, P Gangamohan, Suryakanth V. Gangashetty, and B. Yegnanarayana. Analysis of excitation source features of speech for emotion recognition. In *INTERSPEECH*, pages 1324–1328, 2015.

[10] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Analysis of emotional speech at subsegmental level. In *INTERSPEECH*, pages 1916–1920, 2013.

[11] Sudarsana Reddy Kadiri and B. Yegnanarayana. Speech polarity detection using strength of impulse-like excitation extracted from speech epochs. *accepted for publication in ICASSP*, 2017.

[12] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. *Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions*, chapter Analysis of Emotional Speech—A Review, pages 205–238. Springer International Publishing, 2016.

[13] G. Aneeja and B. Yegnanarayana. Single frequency filtering approach for discriminating speech and nonspeech. *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, 23(4):705–717, April 2015.

[14] Vishala Pannala, G. Aneeja, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Robust estimation of fundamental frequency using single frequency filtering approach. In *INTERSPEECH*, pages 2155–2159, 2016.

[15] B. Yegnanarayana, BHVS Narayana Murthy, and Sudarsana Reddy Kadiri. Single frequency filtering for time delay estimation from multispeaker data. *submitted to IEEE Signal Process. Letters*, 2017.

[16] Nivedita Chennupati, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Intelligibility improvement of speech in noise using single frequency filtering approach. *Speech Communication (under revision)*, 2017.

[17] Nivedita Chennupati, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Significance of phase in single frequency filtering for reconstruction of speech signals. *submitted to IEEE Signal Process. Letters*, 2017.