Computers and Electrical Engineering 55 (2016) 12–23

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Classification of speech under stress using harmonic peak to energy ratio

Suman Deb*, S. Dandapat

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 21 September 2016
Accepted 22 September 2016

Keywords:
Harmonic peak
Signal energy
Speech under stress
Binary-cascade

ABSTRACT

This paper explores the analysis and classification of speech under stress using a new feature, harmonic peak to energy ratio (HPER). The HPER feature is computed from the Fourier spectra of speech signal. The harmonic amplitudes are closely related to breathiness levels of speech. These breathiness levels may be different for different stress conditions. The statistical analysis shows that the proposed HPER feature is useful in characterization of various stress classes. Support Vector Machine (SVM) classifier with binary cascade strategy is used to evaluate the performance of the HPER feature using simulated stressed speech database (SSD). The performance results show that the HPER feature successfully characterizes different stress conditions. The performance of the HPER feature is compared with the mel frequency cepstral coefficients (MFCC), the Linear prediction coefficients (LPC) and the Teager-Energy-Operator (TEO) based Critical Band TEO Autocorrelation Envelope (TEO-CB-Auto-Env) features. The proposed HPER feature outperforms the MFCC, LPC and TEO-CB-Auto-Env features. The combination of the HPER feature with the MFCC feature further increases the system performance.

1. Introduction

Speech under stress is defined as the speech produced under any stress conditions, which perturbs speech production from the neutral condition. There are several reasons that cause stress. Some of the reasons are workload, glottal abnormalities, task demand, noisy environment (Lombard effect), specific emotions such as sad, angry and anxiety [1]. This may results in alteration of speech production mechanism from the neutral condition. Due to this, the performance of speech recognition or speaker recognition decreases under stress conditions. Analysis of speech under stress may improve the performance of speech recognition or speaker recognition. Therefore, analysis of speech under stress is very useful for man-machine interaction.

Analysis of speech under stress can be divided into two parts, feature extraction part and modeling/classification part. In feature extraction part, the desired information is extracted. The modeling/classification part consists of two stages, the training stage and the testing stage. During training, the parameters of the model are updated. In the testing stage, a score is calculated and based on that we make a classification decision. Various modeling techniques have been used for classification of speech under stress. Hidden Markov Model (HMM) [2,3], Artificial Neural Network (ANN) [3] and Support Vector Machine (SVM) [4,5] are used extensively. The performance of speech under stress classification depends on the model

* Corresponding author.
E-mail addresses: suman.2013@iitg.ernet.in (S. Deb), samaren@iitg.ernet.in (S. Dandapat).

http://dx.doi.org/10.1016/j.compeleceng.2016.09.027
0045-7906/

chosen as well as the type of the feature used. Researchers have done many experiments in this regard. First, continuous features including energy, timing and pitch related features provide important cues about various stress conditions [1]. The mel frequency cepstral coefficients (MFCC) feature capturing vocal tract information has been used for speech under stress classification [6,7]. The Linear prediction coefficients (LPC), derived from the linear source filtering concept, is tested for classification of speech under stress [7]. Zhou et al. have shown that Teager-Energy-Operator (TEO) based Critical Band TEO Autocorrelation Envelope (TEO-CB-Auto-Env) feature, derived from non-linear vortex flow through the vocal tract, successfully classify the speech under different stress conditions [2]. Zao et al. have used the pH time frequency feature for speech under stress classification [8]. It is found that pH feature successfully characterizes different stress conditions. Yao et al. have shown that the features, derived from the physical model, are effective in stress classification [9]. Searching for new feature is always a pivot part in classification of speech under stress.

In this work, we have proposed a new feature, harmonic peak to energy ratio (HPER), for classification of speech under stress. The harmonic amplitude is the index, which measures the breathiness level [10]. The breathiness of speech has been used extensively for the detection of different pathologies from the speech signal [10,11]. It is expected that different stress classes may have different breathiness levels. This gives us motivation to propose the HPER feature for analysis and classification of speech under stress. The major contribution of this paper is proposing a new feature, HPER, for speech under stress classification. Along with this, the other contributions are *i*) binary-cascade multi-classification approach using SVM classifier for SSD database and *ii*) a combination of the HPER and the MFCC features for further analysis of classification performance.

The organization of the paper is as follows. The analysis of the HPER feature for speech under stress classification is explained in Section 2. Section 3 discusses the performance of the HPER feature using SVM classifier, and the conclusion is made in Section 4.

2. Harmonic peak to energy ratio (HPER) for classification of speech under stress

This section discusses the process to compute the proposed feature, harmonic peak to energy ratio (HPER) (Section 2.1), the significance of the HPER feature using statistical analysis (Section 2.2), the details about the database (Section 2.3), the SVM classifier (Section 2.4) and the binary-cascade multi-class classification approach of stress classification (Section 2.5). Harmonic peak to energy ratio (HPER) is the amplitudes of the harmonics relative to the total energy of the speech signal. The energy distributions of different stress classes vary with frequency bands. The high-activation stress classes, like happiness and anger, are more concentrated around high frequency regions. On the other hand, low-arousal stress classes, such as sadness and boredom, are low-pitched signals [8]. Lower order harmonics correspond to the lower frequency components, where as higher order harmonics correspond to the higher frequency components. Therefore, amplitude of different harmonic capture the energy concentration at different frequency components. The HPER measures how the harmonic intensity (energy) varies with respect to the total energy. The harmonic amplitude of speech spectrum has also been used for analysis of breathy voice quality. The breathiness level may also be different for different stress conditions.

2.1. Harmonic peak to energy ratio (HPER)

Harmonic peak to energy ratio (HPER) is defined as a ratio of harmonic peaks to the total energy of the speech signal. The HPER feature vector consists of $HPER_i$ elements, $\mathbf{HPER} = [\mathbf{HPER_1}, \mathbf{HPER_2}, \ldots, \mathbf{HPER_M}]^{\mathbf{T}}$, where $i = 1, 2, \ldots, M$ and M is the number of harmonics. The steps of the proposed feature extraction method are described as follows

i) The speech signal is decomposed into a number of frames of 20 ms length with 10 ms frame shift.
ii) Each frame is multiplied with a hamming window to reduce the signal discontinuities at both the ends.
iii) The pitch frequency for each frame is calculated using autocorrelation method. The complete step of pitch estimation is as follows: For a given signal $x(n)$, the autocorrelation $R_x(m)$ is defined as [12]

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m). \tag{1}$$

Thus, if the signal $x(n)$ is periodic with period T, the autocorrelation is also periodic i.e. $R_x(m) = R_x(m+T)$. For a non-stationary signal like speech, the autocorrelation is defined on short segments of speech signal, and it is given by [12]

$$R_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [x(n+l)w(n)][x(n+l+m)w(n+m)] \tag{2}$$

where $0 \le m \le M_0 - 1$, N represents the section length being analyzed, N' is the total number of samples used in $R_l(m)$ computation, $w(n)$ is the hamming window, l represents the starting sample index of the frame and M_0 represents the total number of autocorrelation points. In pitch estimation, N' is normally set as $N' = N - m$, so that only N samples of the frame $(x(l), x(l+1), \ldots, x(l+N-1))$ are used for autocorrelation estimation. From the autocorrelation