Speech Communication 90 (2017) 1–14

Contents lists available at ScienceDirect

Speech Communication

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

Fourier model based features for analysis and classification of out-of-breath speech

Suman Deb*, S. Dandapat

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

ARTICLE INFO

Article history:
Received 6 October 2016
Revised 17 March 2017
Accepted 13 April 2017
Available online 18 April 2017

Keywords:
Out-of-breath speech
Fourier model
Amplitude difference
Frequency ratio
Mutual information

ABSTRACT

This paper presents a new method of feature extraction using Fourier model for analysis of out-of-breath speech. The proposed feature is evaluated using mutual information (MI) on the difference and ratio values of the Fourier parameters, amplitude and frequency. The difference and ratio are calculated between two contiguous values of the Fourier parameters. To analyze the out-of-breath speech, a new stressed speech database, named out-of-breath speech (OBS) database, is created. The database contains three classes of speech, out-of-breath speech, low out-of-breath speech and normal speech. The effectiveness of the proposed features is evaluated with the statistical analysis. The proposed features not only differentiate the normal speech and the out-of-breath speech, but also can discriminate different breath emission levels of speech. Hidden Markov model (HMM) and support vector machine (SVM) are used to evaluate the performance of the proposed features using the OBS database. For multi-class classification problem, SVM classifier is used with binary cascade approach. The performance of the proposed features is compared with the breathiness feature, the mel frequency cepstral coefficient (MFCC) feature and the Teager energy operator (TEO) based critical band TEO autocorrelation envelope (TEO-CB-Auto-Env) feature. The proposed feature outperforms the breathiness feature, the MFCC feature and the TEO-CB-Auto-Env feature.

1. Introduction

Stressed speech or speech under stress is produced due to any alteration of speech production system from that of the normal condition. The characteristics of speech signal changes under stress conditions. Due to this, the performance of machine is affected in case of human-machine interaction. The causes of stress can be due to the specific emotion, sleep deprivation, perceived threat, glottal abnormalities, workload, noisy environments (Lombard effect) (Hansen and Patil, 2007b). It is expected that different stress classes may have different breath emission levels during speech production. Evaluation of breath emission levels of speech can help improve the recognition of stressed speech, which can further enhance the performance of speech recognition, speaker identification, speaker verification. As a result, effectiveness of the man-machine interaction can increase (Calvo and D'Mello, 2010; El Ayadi et al., 2011).

The recognition/classification system for speech under stress has two stages, feature extraction stage and classification stage (Ramamohan and Dandapat, 2006). Number of methods, including hidden Markov model (HMM), Gaussian mixture model (GMM) and support vector machine (SVM), have been used extensively for stressed speech classification (Wu and Liang, 2011; Mower et al., 2011; Attabi and Dumouchel, 2013; Womack and Hansen, 1999; Kotti and Paternò, 2012). The performance of recognition system depends mostly on the type of feature extracted. Researchers have been performed a lot of experiments in this regard. First, continuous features, that include pitch, formant, timing and energy related features, deliver significant cues about different stress conditions (Hansen and Patil, 2007a). Various spectral features including mel frequency cepstral coefficients (MFCC) (Bou-Ghazale and Hansen, 2000; Sezgin et al., 2012) and linear prediction coefficients (LPC) (Sezgin et al., 2012) play important role for stressed speech analysis. The LP model gives equal importance to the spectra of all frequency bands. This property may not hold good with human hearing as the spectral resolution of human hearing decreases behind approximately 800 Hz (Bou-Ghazale and Hansen, 2000). Perceptual linear prediction coefficients (PLP) and revised perceptual linear prediction coefficients (RPLP) are used for classification of speech under stress (Kaminska et al., 2013). The non-linear Teager energy operator (TEO) based features are used for speech under stress classification and it is shown that they outperform the MFCC

* 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.specom.2017.04.002
0167-6393/© 2017 Elsevier B.V. All rights reserved.

features (Zhou et al., 2001). It is shown that the pH vocal source (pH time frequency) features derived from LP residual signal successfully classify different stress conditions (Zao et al., 2014). The pH feature is a vector of Hurst (H) parameters obtained by applying a wavelet-based multidimensional estimator to the windowed short-time segments of speech.

Despite these features, further study is needed regarding the quality of the voice in delivering stress. Search for new feature is always an important area of investigation in pattern recognition/classification tasks of stressed speech. Acoustic analysis of music suggest that the unique quality of each instrument is due to the unique sequence of the harmony structure, because harmony structure of an interval or chord produces a negative or positive impression on listening (Helmholtz and Ellis, 2009). In this work, we propose a set of features, derived from the harmonic sequences of the Fourier parameters, for out-of-breath speech classification. It is expected that the harmony structure of out-of-breath speech may be different from that of the normal speech. This is the motivation for evaluation of out-of-breath speech using Fourier model.

1.1. Existing similar databases

A number of stressed speech databases have been studied in the literature. The speech under simulated and actual stress (SUSAS) database has been used extensively (Hansen, 1989, 1996; Hansen et al., 1997; Zhou et al., 2001). The SUSAS database is partitioned into five domains: (*i*) talking style, (*ii*) single tracking task or speech produced in noise (Lombard effect), (*iii*) dual tracking task, (*iv*) actual speech under stress and (*v*) psychiatric analysis data. Approximately half of the database contains styled data (such as angry, normal, soft, slow, loud , clear, fast) and Lombard speech. Lombard speech was created by adding 85 dB SPL pink noise through headphones. The data were recorded for 35 aircraft communication words, which contain both mono-syllabic and multi-syllabic words. In Ramamohan and Dandapat (2006), two databases were used, one is the English language database and the other one is the Telugu language (an Indian language) database. Both the databases contain three stress classes, anger, happiness and compassion. The speech sentences, corresponding to Telugu and English, contain seven and five vowels respectively. Eight native Telugu speakers participated in data recording. Shukla et al. (2011), created a simulated stressed speech database. The database is of Hindi language (an Indian language). The database consists of four stress conditions, angry, sad, Lombard and neutral. Fifteen non-professional speakers participated in the recording. A total of 35 isolated key words were selected for the data recordings. These keywords are semantically balanced. EMODB database (Burkhardt et al., 2005) is one of the most popular databases for stress/emotion classification. The EMODB database is recorded for ten German sentences and ten speakers participated for data recording. The database contains seven emotions, anger, anxiety, boredom, disgust, happiness, neutral and sadness. Danish emotional speech (DES) database (Engberg and Hansen, 1996) is recorded for two words, nine sentences and two passages of fluent speech, and four professional actors participated in data recording. The DES database contains five emotion classes, anger, sadness, happiness, surprise and neutral. FAU AIBO database (Steidl, 2009) is a database of spontaneous recording of German children, and it consists of 11 classes. Belfast Naturalistic database (Douglas-Cowie et al., 2000), constructed by the Queen's University Belfast, contains two kinds of recordings: one in the studio and the other one directly from TV programs. The first part (studio recording) is from student conversion and the second recording is an audio-visual recording. VAM database (Grimm et al., 2008) is an audio-visual emotional database, and the recording is done with German TV talk show. The database contains spontaneous speech recorded from unscripted discussion by the guests of the TV show. SmartKom multimodel corpus (Schiel et al., 2002), created by the Institute for Phonetik and Oral Communication in Munich, contains multiple audio channels and two video channels. The database contains Wizard-Of-Oz dialogues in English and German. The corpus was developed for gesture and voice recognition module for man-machine interfaces. Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso et al., 2008), recorded by the University of Southern California, is the English language database, and ten speakers participated in the data recording. The IEMOCAP database is recorded during spontaneous spoken communication environments.

1.2. Shortcomings of the existing databases and contributions of this article

The databases, discussed above, contains simulated stress classes. All these databases can be categorized using the descriptor valence-activation (Kotti and释Paternò, 2012). The valence is a measure of pleasure associated with stress condition and it ranges from positive to negative. The activation represents how dynamic the stress or emotional state is. In this work, a new stressed speech database is created, which contains three classes of stressed speech, out-of-breath speech, low out-of-breath speech and normal speech. Here, stress is induced by physical exercise, which perturbs the breath emission levels. The categorization of the recorded database is based on the descriptor breath-emission. The out-of-breath speech is defined as the speech produced with excessive emission of breath. How efficiently breath emission level returns to normal after exercise would be based on the person's physical fitness. We have also recorded the person's pulse rates under three conditions. This may also help us to detect the person's heart or lung health from the speech signal. It is also expected that different stress classes may have different breath emission levels during speech production, e.g., anger class may have higher breath emission level compared to the sadness class. Therefore, analysis of breath emission levels of speech can help improve the recognition of stressed speech, which can further enhance the performance of speech recognition and speaker identification/verification. The major contributions of the present work include: (*i*) recording a new stressed speech database, which contains out-of-breath speech, low out-of-breath speech and normal speech, and (*ii*) feature extraction using mutual information (MI) on the difference and ratio values of the Fourier parameters.

The organization of the present work is as follows: The details about the database are explained in Section 2. The proposed method of feature estimation using Fourier model is presented in Section 3. Section 4 discusses about the details of the classification method. Section 5 discusses the statistical significance of the proposed features. The performance of the proposed features is evaluated in Section 6, and finally conclusion is drawn in Section 7.

2. Out-of-breath speech (OBS) database

In this work, a new stressed speech database is recorded. This recorded database is named as out-of-breath speech (OBS) database. The database contains three classes of speech corresponding to three different levels of breath emission. These three classes are out-of-breath speech, low out-of-breath speech and normal speech. The out-of-breath speech is defined as the speech produced with excessive emission of breath, where as low out-of-breath speech contains lower level of breath emission compared to the out-of-breath speech but higher than the normal speech. Ten male speakers participated in the recording of the OBS database. The speakers are research scholars of Indian Institute of Technology Guwahati (India), and are knowledgeable and familiar with