ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

A Survey on Gender-Driven Emotion Recognition through Speech Signals for Ambient Intelligence Applications

Rushikesh Gade[1], S.R.Ghulhane[2]

PG Student, Department of E&TC, D.Y.Patil College of Engineering, Talegaon Savitribai Phule University of Pune, Pune, India[1]

Professor, Department of E&TC, D.Y.Patil College of Engineering, Talegaon, Savitribai Phule University of Pune, Pune, India [2]

ABSTRACT: This paper proposes a system that permits recognizing a person's emotional state beginning from audio signal registrations. As of late, studies have been performed on harmony features for speech emotion recognition. It is found in our review that the first- and second-order differences of harmony features also play an important role in speech emotion recognition. Therefore, we propose a new Fourier parameter model utilizing the perceptual content of voice quality and the first- and second-order differences for speaker-independent speech emotion recognition. The new feature extraction algorithm called Power Normalized Cepstral Coefficients (PNCC) that is motivated by sound processing. Major new features of PNCC processing include the utilization of a power-law nonlinearity that replaces the conventional log nonlinearity utilized as a part of MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and a module that accomplishes temporal masking. We additionally propose the utilization of medium-time power analysis, in which environmental parameters are estimated over a longer duration than is usually used for speech, as well as frequency smoothing. The obtained results show also that the features selection adoption assures a satisfying recognition rate and allows diminishing the employed features. Future improvements of the proposed solution may include the implementation of this system over mobile devices such as smartphones.

I.INTRODUCTION

SPEECH emotion recognition, which is characterized as extracting the emotional states of a speaker from his or her speech, is attracting more attention. It is trusted that speech emotion recognition can enhance the performance of speech recognition systems and is subsequently extremely accommodating for criminal examination, intelligent assistance, surveillance and detection of potentially hazardous events, and medicinal services systems. Speech emotion recognition is especially useful in man-machine interaction. A speech recognition system consists of two processes: 1) feature extraction and 2) classification [1]. The feature extraction process picks the characteristics of a sound frame, and a word is chosen in the classification process by analyzing the extracted features. This short for the most part concentrates on the hardware design of feature extraction. The most broadly known feature extraction depends on the mel-frequency cepstrum coefficients (MFCCs), as MFCC-based systems are typically connected with high recognition accuracy. MFCC extraction was implemented with an optimized recognition program running on a low-power reduced instruction set PC processor platform. To diminish energy utilization further, dedicated architectures have been proposed and constructed with fixed-point operations. The previous architectures, however, have not completely considered the arithmetic property of the MFCC extraction algorithm.

To adequately recognize feelings or emotions from speech signals, the intrinsic Features must be extracted from raw speech data and transformed into proper formats that are suitable for further processing. It is a longstanding

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

challenge in speech emotion recognition to extract efficient speech features. Scientists have performed many reviews. To begin with, it is found that continuous features including pitch related features, formants features, energy-related features, and timing features deliver important emotional cues. In addition to time-dependent acoustic features, various spectral features such as linear predictor coefficients (LPC), linear predictor cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC) play a significant role in speech emotion recognition. Bou-Ghazale and Hansen found that the features based on cepstral analysis, such as LPCC and MFCC, outperform the linear features of LPC in detecting speech feelings.

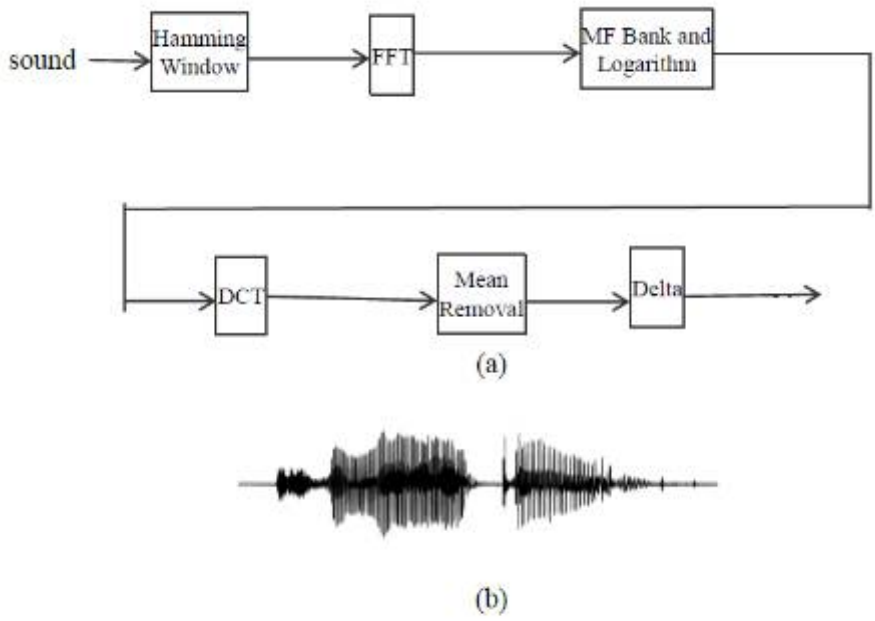

Fig1. (a)Overall process for MFCC features extraction (b)Sound frame.

This concise presents a new energy-efficient architecture for MFCC extraction. Fig1 describe the modified MFCC extraction system and the improvement systems. This concise presents a new energy-efficient architecture for MFCC extraction. Exploring the algorithmic property of MFCC extraction, we renovate the previous architecture with improvement strategies to reduce both hardware complexity and computation time. Subsequently, the energy consumption is remarkably reduced compared with the previous architectures.

II.LITERATURE SURVEY

Yang and Lugger [2] proposed an arrangement of harmony features got from the pitch contour and employed them for the speaker-independent recognition of six classes of emotions: happiness, boredom, neutrality, sadness, anger, and uneasiness. Ruvolo et al. [3] made utilization of the hierarchical aggregation of features to combine short-, medium- and long-scale features. They employed MFCC and LPCC for speaker-independent experiments. Bitouk et al characterized three classes of phonemes in the utterance, namely, stressed vowels, unstressed vowels and consonants, and further computed the statistics of fundamental frequency, first formant, voice intensity, jitter, shimmer and the relative duration of voiced segments for speaker independent experiments. Kotti and Paternò [4] extracted 2,327

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

features in total for speaker-independent recognition that were related to the statistics of pitch, formants, and energy contours as well as spectrum, cepstrum, autocorrelation, voice quality, jitter, shimmer and others. In [5], a small feature set of segment-level features such as the fundamental frequency (F0) and mel-frequency cepstral coefficients (MFCC) with global statistics is used. Many of the present systems developed for automatic speech recognition, speaker identification, and related tasks are depend on variations of one of two types of features: mel frequencycepstral coefficients (MFCC) [6] or perceptual linear prediction (PLP) coefficients [7]. Spectro-temporal features have additionally been as of late presented with promising results (e.g. [8][9]). It has been observed that two-dimensional Gabor filters give a reasonable approximation to the spectro temporal response fields of neurons in the auditory cortex, which has lead to different ways to extract features for speech recognition (e.g. [10], [11]). Shi et al. [12] analyzed the impacts of uncertainty in the phase of speech signals on the word recognition error rate of human audience members. Their results demonstrate that a small amount of phase error or uncertainty does not affect the recognition rate, but a large amount of phase uncertainty has a significant effect on the human speech recognition rate. Hence, the phase may also be important in automatic speech/speaker recognition. Several studies have invested great effort in modeling and incorporating the phase into the speaker recognition process [13]–[14]. The complementary nature of speaker-specific information in the residual phase compared with the information in the conventional MFCCs was exhibited in [14].This paper is organized as follows. Section III presents the Fourier parameter Feature Emotion Analysis based on Fourier series. Section IV discusses the speaker independent recognition. Concluding remarks are in section VI.

III. FOURIER PARAMETER FEATURES FOR SPEECH EMOTION ANALYSIS

In this section, a new model is proposed, with special consideration on three speech emotion databases in two different languages, to extract FP features.

A) Emotion Databases

Three databases are viewed: a German emotional corpus (EMODB), a Chinese emotional
database (CASIA) and a Chinese elderly emotional speech database (EESDB) [15], which are discussed as follows. EMODB was collected by the Institute of Communication Science at the Technical University of Berlin. It has been used by many researchers as a standard data set for examining speech emotion recognition. EMODB comprises 10 sentences that cover seven classes of emotion from everyday communication, namely, anger, fear, joy, sadness, appall, boredom and neutral.

B) Fourier Parameter Features

Harmonics include frequency, amplitude and phase. It has been reported that harmonic frequency features are effective for speech emotion recognition. In this study, we additionally make utilization of harmonic amplitude and phase features. For every frame, FP is evaluated by Fourier analysis. The new speech feature vector Hk might be evaluated for all frames in the speech signal from 1 to ℓ (number of frames). The averaged H_3 among different emotions for one person. It is observed that amplitudes vary with various classes of feelings. We likewise observe the mean of each phase of speech with different emotions, yet the difference is unimportant.

C) Global Fourier Parameter Features

It has been reported that global features are unrivaled in terms of classification accuracy and
computational efficiency. Along these lines, the mean, maximum, minimum, median and standard deviation of the amplitudes of the initial 20 Fourier parameters are computed. The means of H to H are different with regard to seven emotions. The average values of H for sadness, boredom and impartial are higher than disgust, joy and nervousness. The peak of each feeling is at the lower harmonics. The means of six emotions from CASIA.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

IV. SPEAKER INDEPENDENT RECOGNITION

Speaker-independent emotion recognition is one of the most recent challenges in the field of speech emotion recognition. It can adapt to cope with unknown speakers and thus has better generalization than those speaker-dependent approaches. Until now, there have been many reviews gave on speaker-independent emotion recognition. Both MFCC and FP features are extracted for speaker-independent emotion recognition. Continuous features such as fundamental frequency

(F0), energy and zero-crossing rate (ZCR) are also extracted. In pre-processing the acoustic sound pressure wave is converted into a digital signal, which is suitable for voice processing. A microphone can be used to convert the acoustic wave into an analog signal. This analog signal is passed through antialiasing filter to compensate for any channel impairments. The antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate before going through the process of sampling. Then the analog signal is passed through an A/D converter. Today's A/D converters for speech applications typically sample with 16 bits of resolution with 8000–16000 samples per second. The speech is further windowed by passing it through a Hamming window and a suitable frame size of 10 to 30 m sec is chosen.

1. MFCC Extraction Features

MFCC was initially introduced and applied to speech recognition in [16]. It has been prominently utilized for speech emotion recognition [17]. By considering the reaction of human ears to various frequencies, the Mel frequency is determined according to the characteristics of human audition. In this study, MFCC features were extracted for comparison with the proposed FP features. For emotion recognition, MFCC features usually include mean, maximum, minimum, median, and standard deviation. All speech signals were initially filtered by a high-pass filter with a pre-emphasis coefficient of 0.97. The first 13 MFCCs and the associated delta- and double-delta MFCCs were extracted to form a 39-dimensional feature vector. Its mean, maximum, minimum, median and standard deviation were further derived out, which led to a 195-dimensional MFCC feature vector in total.

2. Fourier Parameter Features

We extracted a arrangement of FP features from speech signals as depicted in Sections Fourier Parameter Features and Global Fourier Parameter Features. Here, the initial 120 harmonic coefficients were extracted. The dynamic features were extracted so that temporal derivative features may enhance the performance of emotion recognition. As such, the FP feature vector is comprised of amplitude (H), first-order difference (Δ H) and second-order difference (ΔΔH). Their minimum, maximum, mean, median and standard deviation were also computed. There were a total of 1,800 features for speaker-independent speech emotion recognition. Feature Normalization is an important aspect for a robust emotion recognition system.

3.Pitch Pattern Feature

The prosody of synthetic speech is generally not the same as natural speech and therefore the pitch pattern is another good candidate feature for a countermeasure. The pitch pattern, Φ[n,m], is calculated by dividing the short-range autocorrelation function, r[n, m] by a normalization function, p[n, m] which is proportional to the frame energy

$$\Phi[n, m] = \frac{r[n, m]}{p[n, m]},$$

where

$$r[n, m] = \sum_{k=-N/2}^{N/2} x[n+k-N/2]\, x[n+k+N/2]$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

$$= \sum \quad p[n, m] \quad {}^{1}_{2}{}^{2}[+-/2] + {}^{1}\sum_{2}{}^{2}[++/2$$

and n, m are the sample instant and lag, respectively, over which the autocorrelation is computed.

The lag parameter is chosen such that pitch frequencies can be observed [72]; in this work, we choose $32 \leq m \leq 320$ for a sample rate of 16kHz. Once the pitch pattern is computed, we segment it into a binary pitch pattern image through the rule

$\Phi seg\ [n,m] = \sum 1\ \Phi[n,m] \geq \Theta;\ 0\ \Phi[n,m] < \Theta$

where θ is a threshold; we set $\theta = 1/2$ for all n, based on preliminary results on the development set.

4. Classifier

1. Support Vector Machine

SVM is a binary classifier to analyse the data and recognize the pattern for classification. The main goal is to design hyper plane that classifies all the training vectors in different classes. The aim is to

determine a function which obtain the hyper plane. Hyper plane separates two classes of data sets. The linear classifier is defined as the optimal separating hyper plane.

5. Database

Earlier the database used for automatic speech emotion recognition research was based on acted speech and the researcher tends to have real based data. Database is divided into three types.

1. Acted speech: actors are asked to express deliberately the human emotions which are predefined.
2. Real life speech: the natural response to the conversion of human with spontaneous reaction which are authentic in nature. Example: call center.
3. Elicited emotional speech in which the emotions are induced with self-report instead of labelling, where emotions are provoked and self-report is used for labelling control.

V. CONCLUSION

In previous studies, different features were utilized for speech emotion recognition. An energy-efficient MFCC extraction architecture has been introduced for speech recognition. The MFCC extraction algorithm is modified to minimize computation time without degrading the recognition accuracy noticeably. In this paper, we proposed a new FP and MFCC model to extract salient features from emotional speech signals and validated it on three well-known databases including EMODB, CASIA and EESDB. It is observed that different emotions did lead to different Fps. The review demonstrated that FP features are effective in characterizing and recognizing emotions in speech signals. In addition, it is possible to enhance the performance of emotion recognition by combining FP and MFCC features. These results establish that the proposed FP model is useful for speaker-independent speech emotion recognition. Also, the proposed architecture employs floating-point arithmetic operations to minimize the operation bit-width and the total size of LUTs. The adequacy of energy consumption makes the proposed architecture a promising solution for energy-limited speech recognition systems.

REFERENCES

[1] M. Kotti and F. Paternó, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," Int. J. Speech Technol., vol. 15, pp.131–150, 2012.
[2] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Process., vol. 90, pp. 1415–1423, 2010.
[3] P. Ruvolo, I. Fasel, and J. R. Movellan, "A learning approach to hierarchical feature selection and aggregation for audio classification," Pattern Recog. Lett., vol. 31, pp.1535–1542, 2010.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

[4] M. Kotti and F. Paternó, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," Int. J. Speech Technol., vol. 15, pp.131–150, 2012.
[5] S. Ananthakrishnan, A. Vembu, and R. Prasad, "Model-based parametric features for emotion recognition from speech," in in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011, pp. 529–534.
[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366.
[7] H. Hermansky, "Perceptual linear prediction analysis of speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752.
[8] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, March 2010, pp. 4286–4289.
[9] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," Speech Communication, vol. 53, no. 5, pp. 736–752, May-June 2011.
[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in Proc. INTERSPEECH, pp. 1517–1520.
[11] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in Proc. Int. Conf. Acoust., Speech, Signal Process., 1990, pp. 857–860.
[12] G. Shi et al., "On the importance of phase in human speech recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1867–1874.
[13] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," Proc. ICASSP, vol. 1, pp. 133–136.
[14] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC
features for speaker verification," IEEE Signal Process. Lett., vol. 13, no. 1, pp. 52–55.
[15] K. X. Wang, Q. L. Zhang, and S. Y. Liao, "A database of elderly emotional speech," in Proc. Int. Symp. Signal Process. Biomed. Eng Informat. , 2014, pp. 549–553.
[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech Signal Process., vol. 28, no. 4, pp. 357–366.
[17] N. Kamaruddina, A. Wahabb, and C. Quek, "Cultural dependency analysis for understanding speech emotion,"Expert Syst. Appl., vol. 39, pp. 5115–5133, 2012.