ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

Auditory Processing of Speech Signals for Speech Emotion Recognition

Prashant Aher[1], Alice Cheeran[2]

Department of Electrical Engineering, Veermata Jijabai Technological Institute (VJTI), Mumbai, India[1,2]

Abstract: Feature extraction is most crucial in automatic speech emotion recognition (SER). The performance of cepstral features like Mel Frequency Cepstrum coefficient (MFCC) is good in clean environments but degrades when there exists data mismatch between training and testing phase. An Auditory based feature extraction for SER in noisy environment to recognize and classify the speech emotion from Berlin emotional speech database is presented. The proposed model consists of cochlear bandpass filterbank with zero-crossing for frequency estimation. Features extracted from input speech samples are fed to Support Vector Machine (SVM) classifier with RBF kernel function for classification. As shown in our results, in speech emotion recognition task, both MFCC and proposed feature have recognition accuracy of 81.9% and 89% respectively in clean testing conditions but when SNR of testing speech samples drop to 5 dB recognition accuracy of MFCC feature is 11% while proposed feature achieves an accuracy of 25% , which shows noise robustness of proposed features.

Keywords: Cochlea, cochlear filterbank, mel filterbank, noise robustness, RBF kernel, speech emotion recognition, SVM

I. INTRODUCTION

Feature Extraction is the most important task in Speech Emotion Recognition. Any recognition or classification system is said to be successful if front end features carry enough discriminative information. The performance of SER system usually degrades when there is environmental mismatch between training and testing phase of the system. Examples of such mismatch are various background noises that affect the feature extraction. We have proposed an Auditory based feature extraction inspired by the motion of Basilar Membrane in cochlea that maintains the recognition accuracy in presence of noise.

Speech features can be extracted by modelling human voice production or modelling the peripheral auditory systems [1]. Prosodic features and cepstral features can be used for speech emotion recognition. SER using Pitch, Energy, Formant frequency [6], Zero Crossing Rate, Linear Prediction Cepstrum Coefficients (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) [4] have been investigated and achieved variable classification accuracy of 62.35% to 97.8% depending on the combination of features used for classification.

There has been considerable research in the modelling of peripheral auditory systems. Cepstral features such as Mel Frequency Cepstral Coefficients (MFCC) based on Fourier Transform performs well in clean acoustic environments but recognition performance degrades with mismatched condition. The Fourier Transform has fixed time frequency resolution and well defined inverse transform. Despite its simplicity and efficient computational algorithm, when applied in speech processing, time frequency decomposition of Fourier Transform is different from the mechanism in the human auditory systems. Because it generates the pitch harmonics in the entire speech band and its individual frequency bands are in linear distribution, which is different from the non-linear frequency distribution in human cochlea [1].

In Auditory research, many mathematical models have been defined to simulate the travelling waves, auditory filters and the frequency response of Basilar Membrane [8, 10]. The Gammatone Filter bank has been used as a cochlear model to decompose speech signals into the output of number of frequency bands. After passing speech signals through bank of cochlear filters they can be decomposed into number of frequency bands. The frequency distribution of the cochlear filters is similar to the one in cochlea and the impulse response of filter is similar to that of travelling wave.

In this paper, robust feature Cochlear Filterbank Coefficients with zero crossing is introduced for Speech Emotion Recognition from speech signals, even in noisy environment. Berlin Emotional Database is used to train and test SER system and SVM is used as classifier. The paper deals with comparative analysis of recognition rate using cepstral features like MFCC and cochlear features.

The paper is organized as follows: Section II introduces the impact of various speech parameters on emotion variation and different feature extraction methods. Section III presents details of database used, experiment designs and comparative analysis of feature extraction. Conclusions and future work are presented in section IV.

II. FEATURE EXTRACTION

Speech features extracted from speech signal contains a lot of information [7] and the different parameters result in changes in emotion. Thus the most important step in speech emotion recognition is to extract feature parameter, which can express the different emotions of speech [4]. Some common features are speech rate, energy, pitch, formant and cepstral features such as Linear Prediction

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

Coefficients, Linear Prediction Cepstrum Coefficient (LPCC), Mel-Frequency Cepstrum coefficients (MFCC) and its first derivative and so on [5]. When people are in different emotional state, the speech parameters changes the speak rate, pitch, energy, spectrum etc. is shown in table I.

TABLE I

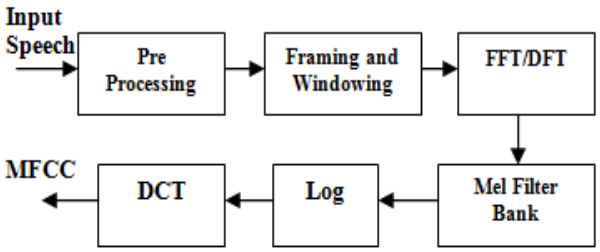
SUMMARY OF THE EFFECTS OF SEVERAL EMOTION STATES ON SELECTED ACOUSTIC FEATURES [4]

Emotion	Pitch			Energy	Spectrum
	Mean	*Variance*	*Variation Range*	*Mean*	*High Frequency Compo-nents*
Anger	Highest	Highest	Increase	Highest	Most
Disgust	Lowest	-	Increase	Lowest	Decrease
Fear	Highest	-	Increase	Normal	Increase
Boredom	Lowest	-	Decrease	Lowest	-
Happiness	Higher	Increase	Increase	Highest	Increase
Sadness	Lower	Decrease	Decrease	Lower	Decrease
Neutral	Normal	Normal	Normal	Normal	Normal

In our work, we extracted MFCC, cochlear features and make use of them to classify the speech emotion.

A. Mel Frequency Cepstrum Coefficients(MFCC)

The Mel Frequency cepstrum is a representation of the short term power spectrum of a voiced signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [9]. It is based on the characteristics of the Human ear's hearing which uses a nonlinear frequency distribution to simulate the human auditory system.



s

Fig. 1 Block Diagram of MFCC Feature Extraction

The mel scale can be calculated as,

$$Mel = 2595\log_{10}\left(1+\frac{f}{700}\right)$$

where f the frequency in Hz and Mel is the perceived frequency in Mels. Using Discrete Fourier transform (DFT) each frame is converted from time domain to frequency domain. Mel frequency scale can be obtained from physical frequency using formula given in (1). DFT Power spectra are weighted by magnitude frequency response of triangular filter bank which are equally spaced using mel scale. Logarithm of output values are calculated to get log energies. In the final stage, the Discrete Cosine Transform (DCT) is used to decorrelate the features. Finally, twelve cepstral coefficients are obtained.

MFCC in the low frequency region has a good frequency resolution and the robustness to noise is also very good, but the high frequency coefficient of accuracy is not satisfactory [4]. So we give up the high-level order of the MFCC and use only low-level order as audio feature parameter.

B. Cochlear Filterbank Cepstral Coefficients

The block diagram of suggested feature extraction using auditory processing of speech signals (i.e. Cochlear filter bank) is depicted in Fig. 2 which consists of bank of band pass filters which simulates the frequency selectivity of Basilar Membrane, common in most auditory models [3] after getting those cochlear coefficients they are forwarded to non-linear stage which performs a series of nonlinear signal processing to simulate the transformation of the mechanical vibration of the basilar membrane into neural firings of auditory nerve fibres.

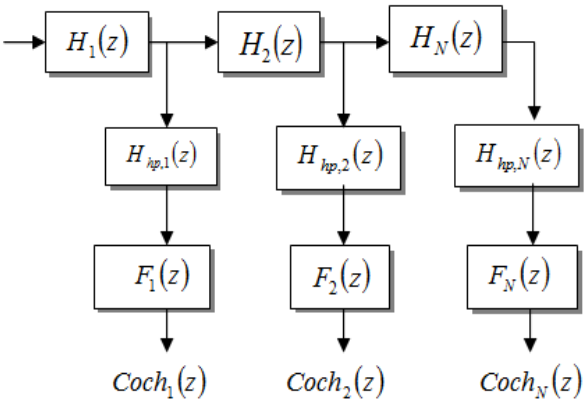


Fig. 2 Block Diagram of Feature Extraction using Cochlear Filters.

Cochlear filters composed of low pass travelling wave filters implemented by a cascade of linear filters simulates the combination of travelling waves progressing through the cochlea. For implementation Gammatone filter bank is used to process audio waveforms which decompose it into number of frequency bands. Output of each filter models the frequency response of basilar membrane at a single place. $H_{hp,i}(z)$ is a one pole high pass filter that models the pressure to velocity transformation and $F_i(z)$ is notch filter which adds notch at one octave below centre frequency by which total response shows two resonance frequency which is similar to biological observations.

Wave propagates from base to the apex of cochlea and high frequency shows maximum excitation near the base while low near the apex. Thus the resonance frequency of $H_N(z)$ decrease as the index N increases. Centre frequencies of Filter bank are distributed in proportion to ERB scale. The transfer function of each cochlear filter coefficient is expressed as,

$$Coch_i(z) = H_{hp,i(z)}F_i(z)\prod_{N=1}^{i} H_N(z)$$

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

An EIH auditory model proposed by Ghitza, a computational efficient and robust for use as front-end speech recognition composed of array of level crossing detectors attached to the output of each cochlear filter [3]. Proper determination of the number of levels and level values is very important for acceptable performance specifically in noisy environments. On the other hand presented model utilizes only zero-crossing for frequency information. The use of zero-crossing in estimating frequency makes it more robust to noise without serious efforts to determine parameters associated with level and shows noise robust property explained by dominant frequency principle.

Hence cochlear filter coefficients are fed to zero crossing detectors. Sign of successive samples are checked for zero crossing. ZC is computed by checking samples in pairs and using function.

$$ZC(m) = \frac{1}{2}\sum_{i=1}^{m} \left| \mathrm{sgn}(c[n]) - \mathrm{sgn}(c[n-1]) \right|$$

where c[n] are filtered samples, 'm' is the filter index and sgn(.) is the sign function returning ± 1 depending on the sign of output sample.

III. EXPERIMENTS

A. Database and Experimental Conditions

The database used in this paper is Berlin Emotional Speech Database which is simulated speech database. It contains seven basic emotions Anger, Boredom, Disgust, Fear, Happiness, Sadness and Neutral recorded from 10 speakers (5 males and 5 female speakers). There are totally about 500 speech samples in which 286 speech samples are of female voice and 207 samples are of male voice. The length of speech sample varies from 2 seconds to 7 seconds and age of speakers varies from 21 to 35 years. All speech samples are single channel data sampled at 16 KHz, 16 bit resolution and end pointed (i.e. there is no initial or final silence) [11].

After evaluation of features from speech samples, Support Vector Machine (SVM) is used for multi-class classification. SVM trained using clean speech samples and testing sets were obtained by mixing clean testing utterances with white Gaussian noise at different SNR level. In total five testing conditions were obtained i.e. noisy speech at 0dB, 5dB, 10dB, 15dB and clean speech. This database was used for the study of noise robustness when training and testing conditions do not match. Support Vector Classifier was trained using clean training set and tested on noisy speech at four SNR level. Speech samples of four different emotions (i.e. Happiness, Anger, Sadness and Boredom) were taken to investigate the efficiency of the proposed Automatic Speech Emotion Recognition.

In SVM we have used Radial Basis Function (RBF) kernel that non-linearly maps the samples to higher dimensional space and is given by,

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right)$$

Where, σ is variance of kernel. Additionally x_i and x_j are support vectors and testing data respectively. Emotions were classified as one vs. rest using five-fold cross validation to improve the accuracy.

Frequency response of the 13 cochlear filters which are used for frequency decomposition of input speech sample are shown in Fig. 3. The magnitude response shows an asymmetric property. Each response shows a long tail on lower frequency side and steep slope at higher frequency side. Also, higher frequency filter shows sharper resonance than lower frequency filter [3].

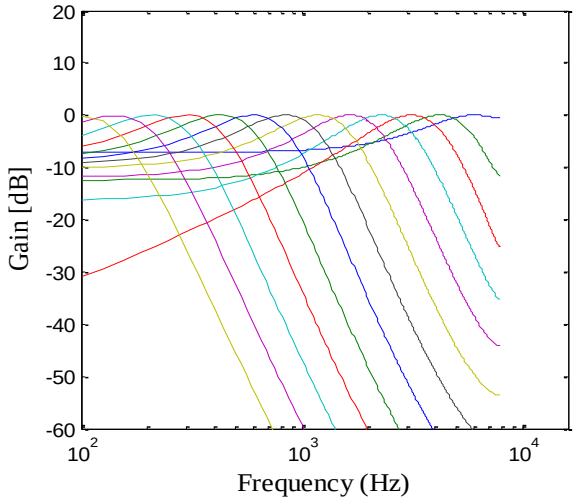


Fig. 3 Frequency Response of 13 Cochlear Filters.

B. Compare MFCC and Cochlear Features

The aim was to find most discriminative feature for speech emotion recognition in presence of noise to improve the overall performance and optimize the feature extraction.

Based on the analytical study, details of MFCC and cochlear filterbank coefficient feature extraction can be summarised as follows: In MFCC 512 point FFT of hamming windowed signal is obtained to convert it from time domain to frequency domain. Log filterbank amplitudes of frequency domain signal were computed in which mel scale is used for frequency distribution. Finally discrete cosine transform is used to decorrelate the features. In cochlear filterbank feature extraction first the cochlear speech samples are passed through the band-pass filter bank. The ERB scale is used for the filterbank distribution. Filterbank coefficients were passed through zero crossing detectors for frequency estimation per channel.

Table II summarizes the SER accuracy of the cochlear filterbank coefficient in comparison with MFCC tested on speech samples with clean and four different SNR levels.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

TABLE II
COMPARISON OF RECOGNITION RATES (%) WITH MFCC AND COCHLEAR FEATURES TESTED IN MISMATCHED CONDITION

Testing SNR	Recognition Accuracy (%)	
	MFCC	Cochlear Features
Clean	81.90	89
15 dB	31.81	50.6
10 dB	20.9	25
5 dB	11	25
0 dB	3.9	8.7

Best cost value 'c' and gamma 'g' for RBF kernel function obtained from cross validation were 0.5 and 4 respectively in support vector machines. Using above optimized values we conduct emotion recognition experiment on the testing set, and the result are shown in Fig. 4. In clean testing condition, the cochlear feature gives comparable result to MFCC. As the noise level in testing data increases, the performance of proposed feature is better than MFCCs performance. For example, when SNR of testing set drops to 5 dB, accuracy of MFCC drops to 11% while cochlear features still achieves 25% accuracy.

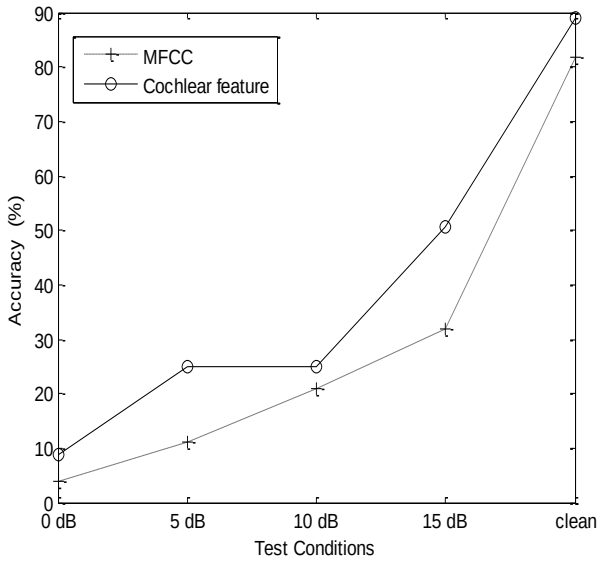


Fig. 4 Comparison of MFCC and proposed cochlear features tested on speech samples with White Gaussian Noise

IV. CONCLUSION

An Auditory based feature for robust speech emotion recognition was presented in this paper. This paper uses only cepstral features for emotion recognition. Our experimental result shows that under various environmental conditions happening in real time situations, new feature gives more recognition accuracy than the MFCC and other prosodic features.

Comparative analysis of recognition rate using different kernel function for non-linear mapping in SVM and evaluation of SER system with the combination of cepstral and prosodic feature to improve recognition accuracy is our future work.

REFERENCES

[1] Qi Li and Yan Huang, "Robust speaker identification using an auditory based feature" ICASSP 2010.
[2] T. seehapoch and S. Wongthanavasu, "Speech emotion recognition using Support Vector Machines", IEEE Int. conference on knowledge and smart technology (KST), 2013.
[3] D. Kim, S. Lee and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real world noisy environments," IEEE Transaction on Speech and Audio Processing, vol. 7, No. 1, January 1999.
[4] P. Shen and X. Chen," Automatic speech emotion recognition using support vector machine", in Proc. 2011 IEEE Int. Conf. on Electronic & mechanical Engg. And Information Technology, pp. 621, August 2011.
[5] T. Pao, C. Wang, and Yu Li, "A study on the search of the most discriminative speech feature in the speaker dependent speech emotion recognition", 2012 fifth Int. symp. on parallel Architectures, Algorithms And Programming.
[6] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification ", in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol.1, pp.593-596, Montreal, May 2004.
[7] S. Dey, R. Ranjan, R. Padmanabhan, and H. Murthy, "Feature Diversity for Emotion , Language, and Speaker Verification ," in 2011 National Conference on Communication s (NCC), Banglore,2011.
[8] J. M. Kates, "A time domain digital cochlea model," IEEE Trans. on Signal Processing, vol. 39, pp. 2573-2592, December 1991.
[9] Hicham Atassi, Anna Esposito, and Zdenek Smekal, "Analysis of high level feature for vocal emotion recognition," in 2011 34th Int. Conf. Telecommunication and signal Processing, Budapest, 2011.
[10] B. C. J. Moor and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," J. Acous. Soc. Am., vol. 74, pp. 750-752, 1983.
[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss," A database of German Emotional Speech " Proc. Interspeech 2005, Lisbon, Portugal, pp. 1517-1520.
[12] J. Manikandan, B. Venkataramani, K. Girish, H. Karthik and V. Sidharth, "Hardware Implementation of Real –Time Speech Recognition System using TMS320C6713 DSP", 24th Annual conference on VLSI Design, IEEE, 2011.
[13] S. Dey, R. Rajan, R. Padmanabhan and H. Murthy, "Feature Diversity for Emotion, Language and Speaker Verification" IEEE, 2011.
[14] L. Rabiner and B. Juang, "Fundamentals of speech recognition", Pearson Education, 2003.
[15] S. Karimi and M. Sedaaghi,"Best Features for Emotional Speech classification in the Presence of Babble Noise", ICEE, Iran, 2012.