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

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

# Comparative Analysis of Speech Features for Speech Emotion Recognition

**Prashant Aher[1], Alice Cheeran[2]**

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

**Abstract:** In this paper, comparative analysis of speech cepstral features is performed to recognise emotion. We identify two effective feature namely, Mel Frequency Cepstral Coefficient (MFCC) and Cochlear Filterbank cepstral coefficients extracted from speech signal. MFCC as a baseline approach is compared to the feature extracted from cochlear filterbank with zero crossing at the output of each channel. Extracted features are fed to Support Vector Machine (SVM) classifier. As shown in our results, cochlear feature provide highest recognition accuracy provided using linear kernel. It gives 89.67% classification accuracy for Berlin Emotional Speech Database. A study on noise robustness of above mentioned feature was also carried out. MFCC and cochlear feature have recognition accuracy of 81.9% and 86% respectively in clean testing conditions with RBF kernel function but when mismatch between training and testing set increases as in real time situations, recognition accuracy of MFCC feature is 11% while cochlear feature gives accuracy of 25%, which shows that cochlear feature is more robust to noise.

**Keywords**: Cochlea, cochlear filterbank, mel filterbank, mel scale, noise robustness, linear polynomial and RBF kernel, speech emotion recognition, Zero Crossing Peak Amplitude, SVM.

## I. INTRODUCTION

Speech is a complex signal contain rich information including text message, speaker identity, intended emotion and so on. Most of the speech system can process studio recorded neutral speech with accuracy. This is due to the difficulty in modelling of emotion present in the utterance [1]. Accurately determining the emotion present in speech has many applications like virtual class room study, determining emotional / psychological health of a person and other activities. Any recognition system is said to be successful if feature extracted from speech signal carry enough information for classification. Number of features and type of feature play a vital role in emotion recognition. We have presented comparative analysis of two cepstral feature extraction method namely Mel frequency Cepstral Coefficients (MFCC) and features based on cochlear filterbank.

Speech features may be extracted from excitation source, vocal tract or prosodic point of view [1]. From literature it can be observed that many research databases were built for speech emotion research such as, Berlin Emotional Database [2], Spanish [3], Chinese [4], Japanese [5] emotional speech database. Many researchers have studied speech emotion recognition. T. Seehapoch and S. Wongthanavasu [6] have investigated features fundamental Frequency, energy, zero crossing rate, Linear Predictive coding and MFCC from short time wavelet signals for SER and achieved highest accuracy of 98%. Subhadeep Dey and associates [7] have shown that some features are best at discriminating some classes and some other features for other class. Experimentations were done on four features and their combination namely MFCC, LPCC, Modified group delay features (MODGDF) and Mel-slope features (fSlope). S. Karimi and M. Sedaaghi [8] carried out research work to find out features which have best performance in the presence of babble noise and investigated features using sequential forward selection (SFS), sequential backward selection (SBS), sequential floating forward selection (SFFS) and sequential floating backward selection (SFBS) in which SFFS shows best outcomes. Roberto and David used ZCPA model for speech recognition and shown that ZCPA performs better than the MFCC feature in noisy conditions, but degrades in clean condition [9]. Literature survey shows that most speech emotion recognition methods use spectral and prosodic features and using different combination of features lead to quite different recognition rate.

Paper presents two features: the MFCC and Cochlear filterbank coefficient for emotion recognition. The MFCC has been widely used in the related work because Mel frequency is proposed according to the characteristics of the human auditory system [10]. Second approach has advantage over first one in noisy or real time situations because MFCC is based on Fourier transform. Time frequency decomposition of Fourier transform is different from the mechanism in the human auditory systems [11]. Cochlear Filterbank Coefficients with zero crossing is introduced for emotion recognition in noisy environment. The Gammatone Filter bank has been used as a cochlear model to decompose speech signals into the output of number of frequency bands. Speech samples from Berlin Emotional Database are used to train and test SER system SVM is used as a classifier. The paper deals with the comparative analysis of recognition rate using cepstral features, MFCC and cochlear feature using various SVM kernel functions in noisy testing condition to increase efficiency of system.

The paper is organized as follows: Section II introduces MFCC and cochlear filterbank feature extraction. Section III gives mathematical introduction to support vector machines and its various kernel functions. Details of database used, experiment designs and comparative

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

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

analysis of feature extraction is given in section IV. Conclusions and future work are presented in section V.

## II. FEATURE EXTRACTION

Speech emotion recognition is highly dependent on the methods which are adopted for feature extraction. Mel Frequency Cepstral Coefficients (MFCC) and zero crossing obtained after cochlear filter bank processing are the methods used for feature extraction in this section. Speech features extracted from speech signal contains a lot of information [7] and the different parameters result in changes in emotion. Some common features are speech rate, energy, pitch, formant and cepstral features such as Linear Prediction Coefficients, Linear Prediction Cepstrum Coefficient (LPCC), Mel-Frequency Cepstrum coefficients (MFCC) and its first derivative and so on [12]. Figure 1 explains the various building blocks of MFCC.

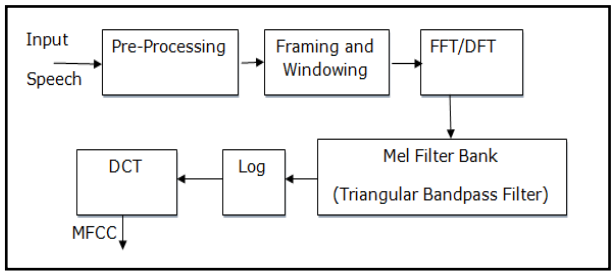


**Fig. 1** Block Diagram of MFCC Feature Extraction

Firstly the input speech signal is pre emphasized using first order FIR filter to spectrally flatten the speech signal using the relation (1).

$$\hat{s}[n] = s[n] - \alpha * s[n-1], where \alpha = 0.9375 \quad (1)$$

It is safe to assume that speech is piecewise stationary. Thus speech signal is framed into 30 to 32 ms frame with an overlap of 20ms. Mathematically framing is equivalent to multiplying signal with sliding window function. It is done using Hamming window function given by,

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), for 0 \le n \le N-1 \quad (2)$$

The pre-emphasized input is split into frames and tshe following windowed output is obtained

$$\hat{x}[n] = w[n]\hat{s}[n], for 0 \le n \le N-1 \quad (3)$$

Next processing step is Fast Fourier Transform, which converts each frame of N samples from time domain to frequency domain. FFT is fast algorithm to implement Discrete Fourier Transform (DFT) which is defined on the N samples and given by,

$$X[k] = \sum_{n=0}^{N-1} \hat{x}[n] e^{-\frac{j2\pi nk}{N}}, for 0 \le k < N-1 \quad (4)$$

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 [13]. It is based on the characteristics of the Human ear's hearing which uses a nonlinear frequency distribution to simulate the human auditory system. The magnitude spectrum $|X[k]|$ is scaled in frequency using the Mel Filter Bank. The Mel Filter bank H(k,m) is collection of triangular filter. The magnitude spectrum $|X[k]|$ is tshen scaled in magnitude by taking the logarithm. The mel scale can be calculated as,

$$Mel = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$

where $f$ the frequency in Hz and Mel is the perceived frequency in Mels. In the final stage, we convert log mel spectrum back to time domain. The result is called Mel Frequency Cepstral Coefficients (MFCC). Cepstral representation of the speech spectrum provide a good representation of the local spectral properties of the signal for given frame analysis because mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to time domain using Discrete Cosine Transform (DCT) [14]. Finally, twelve cepstral coefficients are obtained.

Cochlear Filter Banks are another important category of feature extraction tool which produces peculiar features robust to noisy and contaminated environment. Zero crossings are same during speech production irrespective of the loudness of the utterance [15]. The block diagram of suggested feature extraction using auditory processing of speech signals (i.e. cochlear filter bank) is depicted in Figure 2. Travelling wave filter H(z), Velocity transformation filter T(z) and second filter F(z) are important building blocks of cochlear filter bank.

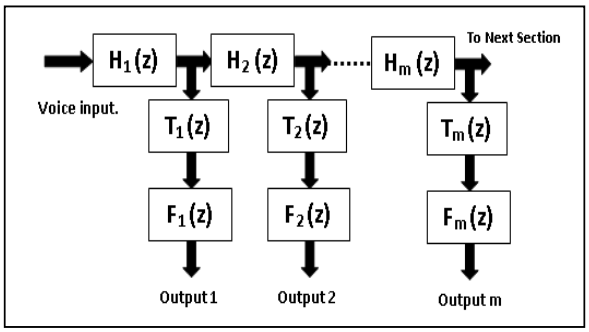


**Fig. 2** Block Diagram of Feature Extraction using Cochlear Filters.

Properly uttered speech segment is passed through travelling wave filter. This captured voice is processed in cascade manner in the various section of the travelling wave filter which possesses the low pass filtering characteristics. Cut off frequency is different for each section [16].

Processed speech signal from travelling wave filter are then passed to the velocity transformation filter for the total dismissal of the low frequency speech signal content. The cut off frequency of filter is kept two octave below the centre frequency of each segment, where each segment is one pole high pass filter. F(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.

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

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

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. 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(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,

$$Coch_i(z) = H(z)F(Z)\prod_{N=1}^{i} H_N(Z) \quad (6)$$

After processing the speech through various sections, zero crossing from each channel is used for recognition purpose. As the zero crossings are not susceptible to the influence of noise hence system proves to be robust in noisy environment. ZC is computed by checking samples in pairs and using function.

$$ZC(m) = \frac{1}{2}\sum_{i=1}^{m}\left|\text{sgn}(c[n]) - \text{sgn}(c[n-1])\right| \quad (7)$$

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

## III. SUPPORT VECTOR MACHINES (SVM) : CLASSIFIER

In regard to the choice of classification method, the kind of application of speech recognition system is crucial. If all patterns in dataset can be separated by straight line or hyperlpane, the dataset is said to be linearly separable. However, there are many problems which are not linearly separable. SVM uses linear models to implement nonlinear class boundaries [14]. It transforms the input space using nonlinear mapping to a new space. Then the linear model constructed in new space can represent a nonlinear decision boundary in the original space. Another component in SVM approach is the maximum margin hyperplane.

The goal of SVM is to produce a model which predicts the target value of test data given only the test data attributes. Given a training set of instance- label pair $(x_i, y_i), i = 1.....l$ where $x_i \in \Re^n$ and $y \in \{-1,1\}^l$, SVM requires the solution of optimization problem,

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l}\xi_i \quad (8)$$

Subject to (1) $y_i\left(w^T\phi(x_i) + b\right) \geq 1 - \xi_i$

(2) $\xi_i \geq 0$.

Here training vector $x_i$ are mapped into higher space by function $\phi$. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space [17]. C > 0 is the penalty parameter of the error term. $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is called kernel function. Optimal value of the "C" assure reliable estimate of the performance of speech recognition method. The kernels like Linear, Polynomial, radial basis function (RBF) and wavelet are often used. Which kernel to be used depends on the specific data and applications. Table I gives examples of most commonly used kernel functions.

TABLE I
SVM KERNELS AND KERNEL FUNCTIONS

| Kernel | Kernel Function |
|---|---|
| Linear | $K(x,z) = x_i^T z_i$ |
| Polynomial | $K(x,z) = (x_i^T z_i + \gamma)^d$, where d is degree of polynomial |
| Radial Basis Function (RBF) | $K(x,z) = \exp(-\frac{\|x_i - z_i\|^2}{2\sigma^2})$ |

## IV. EXPERIMENTS

### A. Database and Experimental Conditions

Berlin emotional speech database (EmoDB) has seven emotions namely, Anger, Boredom, Disgust, Fear, Happiness, Sadness and Neutral. Four emotions happiness, anger, boredom and sadness which are most discriminative are considered for the experimentation. In EmoDB 10 actors, (5 men, 5 women) have participated to create 10 ordinary germen utterances in seven different emotions. More details can be found in [2].

Since the speech input is endpoint detected, windowing is performed on the speech frames. Windowed output is fed to Mel Filterbank for extracting MFCC and each sample results in 13 MFCC coefficients. In case of feature extraction using cochlear filterbank, speech input is passed through 13 channel filterbank for frequency decomposition. Zero crossing is computed at the output of each channel and feature of size $1 \times 13$ is obtained for each input signal.

After evaluation of cepstral features Multi-class SVM is used for emotion recognition. A study of noise robustness of these features is performed in mismatched condition by training the SVM using clean dataset and tested on noisy speech at four SNR levels (i.e. 0dB, 5dB, 10dB and 15dB). Remember that training dataset consist of only clean speech samples while testing dataset consist of noisy as well as clean samples.

This paper uses SVM with a number of kernel functions such as linear, polynomial, wavelet and Radial Basis Functions with N-fold cross validation in experimentation. Cross validation is a common practice used in performance analysis that randomly partitions the data into N complementary subsets, with N-1 of them used for training in each validation and remaining one used for testing [17]. Noise robustness analysis is performed using SVM with RBF kernel that non-linearly maps the samples to higher dimensional space and is given by,

$$K(x,z) = \exp\left(-\frac{\|x_i - z_i\|^2}{2\sigma^2}\right) \quad (9)$$

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

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

Where, $\sigma$ is variance of kernel. Additionally $x_i$ and $z_i$ are support vectors and testing data respectively [6].

### B. Results

In this section results using two approaches namely, MFCC and Cochlear Filterbank Coefficients are evaluated and compared. A recognition performance is studied using several sets of training datasets to decide the optimum number of training samples required for the recognition system and to improve the overall performance.

The recognition accuracy achieved with SVM trained and tested using clean speech samples for MFCC and Cochlear feature input is tabulated in Table II for various SVM kernels.

TABLE II
RECOGNITION PERFORMANCE OF SER SYSTEM FOR DIFFERENT SVM KERNELS IN MATCHED CONDITION

| SVM Kernel | Recognition Accuracy (%) | |
|---|---|---|
| | **MFCC** | **Cochlear Features** |
| **Linear** | 85 | 89.67 |
| **Radial Basis Function (RBF)** | 81.90 | 86. |
| **Wavelet** | 81. | 84 |
| **Polynomial** | 78.10 | 82 |

The experimentation was carried out by varying cost values. By varying the respective parameter, it is found that linear kernel (linear), polynomial kernel at degree 3 (poly3) and RBF kernel at sigma 4 (rbf4) gives best results.

TABLE III
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 |

Comparative analysis of MFCC and Cochlear feature input to study the noise robustness is summarized in Table III in which recognition accuracy is tested using speech samples at four different SNR levels.

Using SVM with 5-fold cross validation best cost value 'C' and gamma 'g' for RBF kernel obtained were 0.5 and 4 respectively.

For performance evaluation, the experiments with all possible combination of different kernel functions and above experimental conditions were performed. It may be observed from table II that MFCC feature with linear kernel function gives maximum recognition accuracy. Table III depict the result of noise robustness analysis with Berlin emotion database and it may be observed that recognition accuracy of MFCC degrades as the SNR level increases while cochlear features provide more recognition accuracy as compare to MFCC.

## V. CONCLUSION

Baseline feature extraction method MFCC is compared to the presented cochlear filterbank cepstral coefficients which composed of cochlear bandpass filter and zero crossing stage at the output of each bandpass filter. In this regards, MFCC features with linear kernel in SVM classifier gives best result for emotion classification and is similar to the results in [17]. Comparative analysis of these two features was carried out to investigate the noise robustness of SER system and it was found that under various environmental conditions happening in real time situations, new feature gives more recognition accuracy than MFCC and other prosodic features.

In this study only cepstral features are considered. Both the cepstral and prosodic feature contains the emotion characteristics and combination of them can be used to increase the recognition accuracy. More work is needed to improve the system so that it can be implemented as real-time SER system.

## REFERENCES

[1] Shashidhar K., Nitin Kumar and K. Rao, "Speech emotion recognition using segmental level prosodic analysis" , IEEE 2011.
[2] 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.
[3] J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, "Analysis and Modeling of emotional speech in spanish ", in proc. ICPhS'99,pp957-960, San Fransisco 1999.
[4] F. Yu, E. Chang and H. Y. Shum, "Emotion Recognition from speech to enrich multimedia content ", in proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp. 550-557, Beijing, China, October 2001.
[5] T. Moriyama, S. Mori and Shinji Ozawa, "A synthesis method of emotional speech using subspace constraints in prosody", Journal of Information processing Society of Japan, pp. 1181-1191,2009.
[6] T. seehapoch and S. Wongthanavasu, "Speech emotion recognition using Support Vector Machines", IEEE Int. conference on knowledge and smart technology (KST), 2013.
[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] S. Karimi and M. Sedaaghi,"Best Features for Emotional Speech classification in the Presence of Babble Noise", ICEE, Iran, 2012.
[9] S. Haque, R. Togneri and David, "An Enhanced Feature Extraction Method for the Zero-Crossing with Peak Amplitude Auditory Model Based on the Mean Discharge Rate", SST, Melbourne, Australia, 2010.
[10] Qingli Zhang, Ning An and Lian Li, "Speech Emotion Recognition using Combination of Featured", Fourth Int. Conf. on Intelligent control and Information Processing, Beijing, China, IEEE 2013.
[11] Prashant Aher and A N Cheeran, "Auditory Processing of Speech Signals for Speech Emotion Recognition", IJARCCE, Vol.3, Issue 5, May 2014.
[12] 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.
[13] 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.
[14] Swapnil Daphal and Sonal Jagtap," DSP Based Improved Speech Recognition System" International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, IEEE, 2011.

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



[15] Chok-Ki Chan, "Speech recognition based on Zero-Crossing rate and energy ", IEEE transaction on Acoustics, Speech and Signal Processing, Vol. 33, No. 10, 1985.
[16] J. M. Kates, "A time domain digital cochlear model," IEEE Trans. on Signal Processing, vol. 39, pp. 2573-2592, December 1991.
[17] http://www.csie.ntu.edu.tw/~cjlin
[18] Qi Li and Yan Huang, "Robust speaker identification using an auditory based feature" ICASSP 2010.
[19] 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.
[20] 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.