Recognition of Human Speech Emotion Using Variants of Mel-Frequency Cepstral Coefficients

Hemanta Kumar Palo, Mahesh Chandra and Mihir Narayan Mohanty

Abstract In this chapter, different variants of Mel-frequency cepstral coefficients (MFCCs) describing human speech emotions are investigated. These features are tested and compared for their robustness in terms of classification accuracy and mean square error. Although MFCC is a reliable feature for speech emotion recognition, it does not consider the temporal dynamics between features which is crucial for such analysis. To address this issue, delta MFCC as its first derivative is extracted for comparison. Due to poor performance of MFCC under noisy condition, both MFCC and delta MFCC features are extracted in wavelet domain in the second phase. Time–frequency characterization of emotions using wavelet analysis and energy or amplitude information using MFCC-based features has enhanced the available information. Wavelet-based MFCCs (WMFCCs) and wavelet-based delta MFCCs (WDMFCCs) outperformed standard MFCCs, delta MFCCs, and wavelets in recognition of Berlin speech emotional utterances. Probabilistic neural network (PNN) has been chosen to model the emotions as the classifier is simple to train, much faster, and allows flexible selection of smoothing parameter than other neural network (NN) models. Highest accuracy of 80.79% has been observed with WDMFCCs as compared to 60.97 and 62.76% with MFCCs and wavelets, respectively.

Keywords Human speech emotion · Mel-frequency cepstral coefficient Probabilistic neural network · Feature extraction · Wavelet analysis

H.K. Palo · M.N. Mohanty (✉)
Department of Electronics and Communication Engineering,
Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India
e-mail: mihir.n.mohanty@gmail.com

H.K. Palo
e-mail: hemantapalo@soauniversity.ac.in

M. Chandra
Department of Electronics and Communication Engineering,
Birla Institute Technology, Ranchi, India
e-mail: shrotriya69@rediffmail.com

A. Konkani et al. (eds.), *Advances in Systems, Control and Automation*,
Lecture Notes in Electrical Engineering 442, https://doi.org/10.1007/978-981-10-4762-6_47

1 Introduction

Currently due to involvement of ambiguous and multi-modal expressive behavior, emotion inference in daily human interaction remains a challenge. Mostly the emotional expressions are in the form of facial or bodily movements, although speech is sole medium of emotional expression during telephone conversation. For such task, detection of emotion arguably remains a challenge. The recognition system demands extraction of suitable features that can best represent speech emotions.

Among different feature extraction techniques such as prosodic, spectral, voice quality, nonlinear, MFCC uses frame-based approach and is widely accepted for speech and emotion analysis [1–3]. Although popular, MFCC feature extraction process does not involve the temporal dynamics among features that is essential for emotional analysis. The drawback can be minimized using derivative and acceleration coefficients of MFCC and has been approached by different researchers [4, 5]. However, MFCC-based features including deltas give information on amplitude and energy, hence unable to provide adequate information on speech emotion. To improve the recognition further, multi-resolution capability of wavelet analysis in extracting MFCC features has been proposed for speaker verification [6]. These literatures suggest for a possibility in enhancement of accuracy if the advantages of both wavelet analysis and differential MFCC are combined. Hence, a novel effort has been made to extract the wavelet-based delta MFCC (WDMFCC) features for comparison purpose.

Most classifiers used in speech analysis are statistical, NN-based, fuzzy logic, and combination of these [1–7]. GMM and HMM have the ability to model the pattern involving large feature sets. Smaller dataset and difficulty in modeling the emotions using conventional statistical classifiers such as GMM and HMM have opened up new avenues for NN-based classifiers [7]. As compared to multilayer perceptron (MLP), PNN is much faster, simpler, easy to implement, and more accurate [8]. Requirement of only one parameter adjustment and absence of any constraint in choosing the parameter with precision makes it superior to RBFN. This has motivated the authors to opt for probabilistic neural network (PNN) classifier in this work.

The organization of the paper is as follows: The feature extraction techniques and the classification model used are explained in Sects. 2 and 3, respectively. A detailed description on the simulation result with a comparison among the state of the art and the proposed features is described in Sect. 4. The conclusion and future research direction are provided in Sect. 5.

2 Feature Extraction Techniques

Initially, standard features based on wavelet analysis and MFCC are extracted and compared for their effectiveness in terms of recognition accuracy and mean square error (MSE). Next to it, delta MFCC features are extracted and compared with the

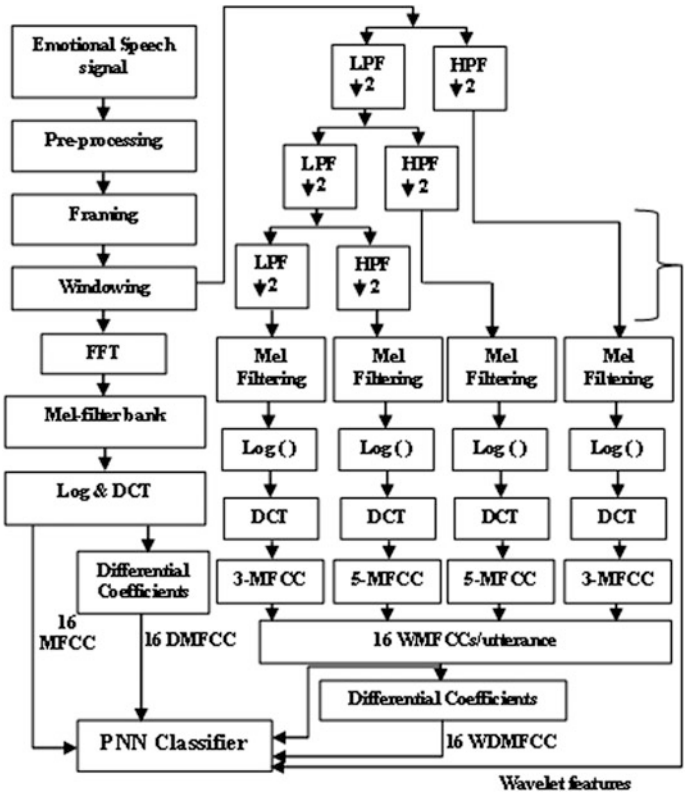

Fig. 1 Proposed feature extraction technique

wavelet-based MFCC and proposed wavelet-based delta MFCC features. The proposed feature extraction technique is shown in Fig. 1.

2.1 MFCC

One of the most dominant and effective cepstrum-based features is MFCC that uses a Mel-scale to wrap the original signal frequency into Mel-frequency. This way both human auditory and hearing mechanism are taken into account since human ear is logarithmic in nature. The relationship used to convert the windowed signal frequency f into Mel-frequency f_m is given by

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

2.2 Wavelet Analysis

In wavelet analysis, the emotional speech signal $s(n)$ is decomposed into approximated and detailed coefficients using a series of low-pass and high-pass filters, respectively [9]. For the low-pass filter having an impulse response $h(n)$, the output can be represented as a convolution and is given by

$$X_L(n) = s(n) * h(n) = \sum_{k=-\infty}^{\infty} s(k)h(n-k) \tag{2}$$

Similarly, the output of the high-pass filter is the convolution of the filter impulse response and the signal under consideration. The sub-sampled output of the filters by a factor of two can be represented as

$$X_L(n) = \sum_{k=-\infty}^{\infty} s(k)h(2n-k) \tag{3}$$

$$X_H(n) = \sum_{k=-\infty}^{\infty} s(k)h(2n+1-k). \tag{4}$$

2.3 MFCC in Wavelet Domain (WMFCC)

The steps of extraction of WMFCC feature are explained below.

(1) Signal decomposition: Initially, decompose the signal into detailed d_i and approximation a_i coefficients at ith level, respectively. A three-level decomposition with Daubechies-4 (dB-4) has been performed here. Emotional utterances of Berlin database has been used for this work having a sampling frequency of 16 kHz. Thus, for 8 kHz bandwidth, the sub-bands are distributed in the range of 0–1, 1–2, 2–4, and 4–8 kHz due to filtering.
(2) The MFCC feature extraction technique as shown in Fig. 1 is applied individually for these sub-bands. Five WMFCCs from each sub-bands of 1–2 and 2–4 kHz and three WMFCCs from each band of 0–1 and 4–8 kHz have been extracted. More WMFCCs are extracted from the middle bands (1–4 kHz) to obtain more perceptual information as the speech signal often lies in this band. The number of WMFCC coefficients obtained this way is reduced to 16 coefficients per utterance.

2.4 Delta MFCC (DMFCC)

The delta MFCC (M_Δ) features of emotional utterances are extracted using the relation:

$$M_\Delta = \beta \times \sum_{q=1}^{2} n \times [\text{MFCC}(r+q) - \text{MFCC}(r-q)], \quad r = 1, 2, \ldots N \tag{5}$$

where MFCC_Δ is the delta features, and to scale the frequency, a value of $\beta = 2$ has been used. N is the number of delta features per utterance, whereas q is the indexing parameter associated with analyzing window.

2.5 Proposed Wavelet Delta MFCC (WDMFCC)

In the proposed technique, the use of wavelet analysis has provided both time and frequency information of the signal, while traditional MFCCs provided the energy or amplitude information. Application of derivatives of MFCC has inducted the dynamic characteristics with the WMFCC features. The steps of extracting WDMFCC features are explained below.

a. Extract the wavelet features as explained earlier using Eqs. (3) and (4) as shown in Fig. 1.
b. Apply the MFCC feature extraction technique to the wavelet features as explained in WMFCC feature extraction technique.
c. Apply the derivative algorithm as given in Eq. (5) to the WMFCC features to extract 16 WDMFCC features per utterance of an emotion.

3 Classification Method

PNN is a nonparametric network having input, pattern, summation, and decision layers as the main constituent sections as shown in Fig. 2.

For any input emotion pattern E with dimension of the vector x and smoothing parameter δ, the output of the pattern layer can be represented as

$$\delta_{u,v}(E) = \frac{1}{(2\pi)^{\frac{x}{2}}\delta^x} \exp\left[-\frac{(E - E_{u,v})^T (E - E_{u,v})}{2\delta^2}\right] \tag{6}$$

where $E_{u,v}$ denotes the neuron vector and is considered to be the center of the kernel function. Here $u = 1, 2, \ldots, U$ is the number of emotional states; $v = 1, 2, \ldots V_u$ and

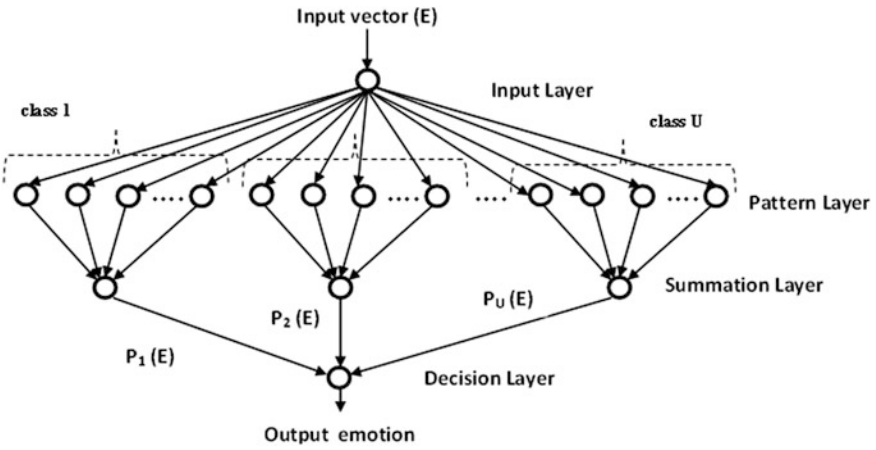

Fig. 2 The structure of PNN

V_u is the total number of feature values in class S_u as given in Eq. (2). Based on the probability distribution function (PDF) of all the neuron, the summation layer summarizes and averages all the neuron outputs of the designated class so as to classify the pattern E to that class S_u.

$$p_u(E) = \frac{1}{(2\pi)^{\frac{x}{2}}\delta^x}\frac{1}{V_u}\sum_{v=1}^{V_u}\exp\left[-\frac{(E-E_{u,v})^T(E-E_{u,v})}{2\delta^2}\right] \tag{7}$$

The decision layer takes a decision on the predicted class using Bayes' approximation of the summation layer neuron output:

$$\widehat{S}(E) = \arg\max\{p_u(E)\} \tag{8}$$

where the computed class of the pattern E is represented by $\widehat{S}(E)$.

4 Results and Discussion

Berlin emotional speech database (EMO-DB) [10] has been chosen to distinguish five emotional states such as angry, anxiety, happy, sad, and neutral. The utterances used for training of the classifier are not involved in testing or validation purpose. Approximately 70% of the data are used for training, and 20% each are used for testing and validation of every emotional state.

The recognition accuracy using the extracted features with PNN classifier is given in Table 1. Comparison of the results has gone in favor of WDMFCC as compared to others. Use of both temporal information due to involvement of

differential values of MFCC and exploration of multi-resolution capability of wavelet analysis is the prime reason of the accuracy improvement.

Wavelet-based features like WMFCC and WDMFCC are better than delta MFCC and MFCC features as observed in Table 1. It is found that involvement of either differential information or wavelet approach in modifying the MFCC extraction technique has enhanced the robustness of the resultant features as compared to the standard MFCC or wavelet features.

A comparison of MSE has provided similar results as shown in Fig. 3. It is found that standard wavelet-based features are more reliable as compared to MFCCs both in terms of classification accuracy and MSE.

To find the response time of the classifier with different MFCC features extracted, the time elapsed both during feature extraction and classification has been compared in Table 2.

Table 1 The recognition accuracy using the extracted features with PNN classifier

Features	Angry (%)	Anxiety (%)	Happy (%)	Neutral (%)	Sad (%)	Average accuracy (%)
MFCC	64.55	62.78	61.44	57.12	58.96	60.97
Wavelets	66.71	64.35	63.42	59.30	60.01	62.76
DMFCC	65.80	64.81	65.54	62.06	61.37	63.92
WMFCC	86.60	84.58	76.21	68.63	66.49	76.50
WDMFCC	89.27	86.64	80.39	75.48	72.15	80.79

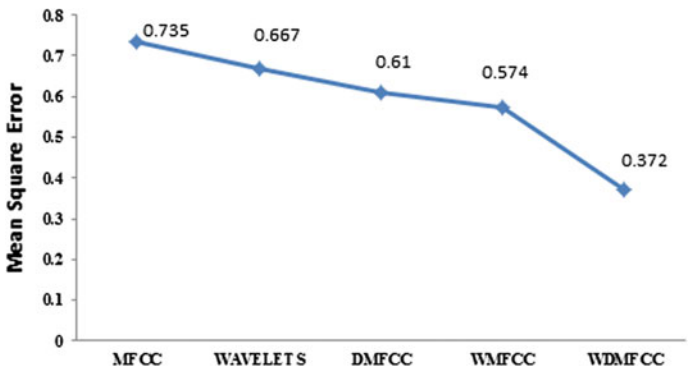

Fig. 3 Comparison of MSE using variants of MFCC with PNN classifier

Table 2 Real-time comparison of different features using PNN classifier for 45 utterances per emotion

Features	MFCC	Wavelets	DMFCC	WMFCC	WDMFCC
Feature extraction time/utterance (s)	0.1159	1.244	0.18809	0.32604	2.19
Classification time/emotion (s)	15.50	13.13	10.56	11.54	13.12
Recognition accuracy (%)	60.97	62.76	63.92	76.50	80.79

5 Conclusion

There is an improvement in feature reliability involving the differential values of MFCC coefficients due to inclusion of temporal emotional information. Application of multi-resolution capability in extraction of MFCC features has resulted more reliable information than standard MFCC as the classification results suggest. Involvement of both wavelet analysis and differential algorithm used for modification of standard MFCC found to be a novel effort in this direction. Other modification techniques that can add valuable emotional information to MFCC features can open up new avenues in the field of emotion recognition.

References

1. Kari, B., Muthulakshmi, S.: Real time implementation of speaker recognition system with MFCC and neural networks on FPGA. Indian J. Sci. Technol. **8**(19), 1–11 (2015)
2. Mohanaprasad, K., Pawani, J.K., Killa, V., Sankarganesh, S.: Real time implementation of speaker verification system. Indian J. Sci. Technol. **8**(24), 1–9 (2015)
3. Subhashree, R., Rathna, G.N.: Speech emotion recognition: performance analysis based on fused algorithms and GMM Modelling. Indian J. Sci. Technol. **9**(11), 1–8 (2016)
4. Mishra, A.N., Chandra, M., Biswas, A., Sharan, S.N.: Robust features for connected Hindi digits recognition. Int. J. Signal Process. Image Process. Pattern Recogn. **4**(2), 79–90 (2011)
5. Kwon, O.W., Chan, K., Hao, J., Lee, T-W.: Emotion recognition by speech signals. In: Interspeech (2003)
6. Kumar, P., Chandra, M.: Hybrid of wavelet and MFCC features for speaker verification. In: 2011 World Congress on Information and Communication Technologies, IEEE, pp. 1150–1154 (2011)
7. Ayadi, E., Kamal, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011). Elsevier
8. Specht, D.: Probabilistic neural networks. Neural Netw. **3**, 109–118 (1990)
9. Palo, H.K., Mohanty, M.N., Chandra, M.: Efficient feature combination techniques for emotional speech classification. Int. J. Speech Technol. **19**, 135–150 (2016)
10. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of the Interspeech, Lissabon, Portugal, pp. 1517–1520 (2005)