Use of Different Features for Emotion Recognition Using MLP Network

H.K. Palo, Mihir Narayana Mohanty and Mahesh Chandra

Abstract Emotion recognition of human being is one of the major challenges in modern complicated world of political and criminal scenario. In this paper, an attempt is taken to recognise two classes of speech emotions as high arousal like angry and surprise and low arousal like sad and bore. Linear prediction coefficients (LPC), linear prediction cepstral coefficient (LPCC), Mel frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP) features are used for emotion recognition using multilayer perception (MLP).Various emotional speech features are extracted from audio channel using above-mentioned features to be used in training and testing. Two hundred utterances from ten subjects were collected based on four emotion categories. One hundred and seventy-five and twenty-five utterances have been considered for training and testing purpose.

Keywords Emotion recognition · MFCC · LPC · PLP · NN · MLP · Radial basis function

1 Introduction

Emotion recognition is a current research field due to its wide range of applications and complex task. It is difficult and challenging to analyse the emotion from speech. Emotion is a medium of expression of one's perspective or his mental

H.K. Palo (✉) · M.N. Mohanty
Siksha 'O' Anusandhan University, Bhubaneswar, India
e-mail: hemantapalo@soauniversity.ac.in

M.N. Mohanty
e-mail: mihirmohanty@soauniversity.ac.in

M. Chandra
Birla Institute of Technology, Ranchi, India
e-mail: shrotriya69@rediffmail.com

I.K. Sethi (ed.), *Computational Vision and Robotics*, Advances in Intelligent Systems and Computing 332, DOI 10.1007/978-81-322-2196-8_2

state to others. Some of the emotions include neutral, anger, surprise, fear, happiness, boredom, disgust, and sadness and can be used as input for human–computer interaction system for efficient recognition. Importance of automatically recognising emotions in human speech has grown with increasing role of spoken language interfaces in this field to make it more efficient. It can also be used for vehicle board system where information of mental state of the driver may be provided to initiate his/her safety though image processing approaches [1]. In automatic remote call centres, it is also used to timely detect customer's emotion.

The most commonly used acoustic features in the literature are LPC features, prosody features like pitch, intensity and speaking rate. Although it seems easy for a human to detect the emotional classes of an audio signal, researchers have shown average score of identifying different emotional classes such as neutral, surprise, happiness, sadness and anger. Emotion recognition is one of the fundamental aspects to build man–machine environment that provides theoretical and experimental basis of the right choice of emotional signal for understanding and expression of emotion. Emotional expressions are continuous because the expression varies smoothly as the expression is changed. The variability of expression can be represented as amplitude, frequency and other parameters. But the emotional state is important in communication between humans and has to be recognised properly.

The paper is organised as follows. Section 1 introduces the importance of this work; Sect. 2 represents the related literature. The proposed method has been explained in Sect. 3. Section 4 discusses the result, and finally Sect. 5 concludes the work.

2 Related Literature

The progress made in the field of emotion recognition from speech signal by various researchers so far is briefed in this section. Voice detection using various statistical methods was described by the authors [2] in their paper. The concept of negative, non-negative emotions from a call centre application was emphasised using a combination of acoustic and language features [3]. A review on various methods of emotion speech detection, features and resources available were elaborately explained in the paper [4, 5]. A tutorial review on linear prediction in speech was explored by the author [6] in his paper, and the algorithm behind the representation of speech signal by LP analysis was suitably explained in [6, 7]. Spectral features such as Mel frequency cepstral coefficient (MFCC) was explained completely in [8, 9]. The concept of linear prediction cepstral coefficient (LPCC) and neural network classifier has been the main focus in this paper [10]. Perceptual linear prediction features of speech signals with their superiority over LPC features are suitably proved with experimental results and algorithms by the authors in [11]. The idea about various conventional classifiers including neural network can be found in the paper of authors [12], while speech emotion recognition by using combinations of C5.0, neural network (NN), and support vector machines (SVM) classification methods are emphasised in [13].

3 Proposed Method for Recognition

Two of the major components of an emotional speech recognition system are feature extraction and classification.

3.1 Feature Extraction

Features represent the characteristics of a human vocal tract and hearing system. As it is a complex system, efficient feature extraction is a challenging task in emotion recognition system. Extracting suitable features is one of the main aspects of the emotion recognition system. Linear prediction coefficients (LPCs) [6, 7] are one of the most used features for both speech and emotional recognition. Basic idea behind the LPC model is that given speech sample at time n, $s(n)$ can be approximated as a linear combination of the past p speech samples. A LP model can be represented mathematically

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

The error signal $e(n)$ is the difference between the input speech and the estimated speech. The filter coefficients a_k are called the LP (linear prediction) coefficients.

One of the most widely used prosodic features for speech emotion is MFCC [8, 9] which outperformed LPC in classification of speech and emotions due to use of Mel frequency which is linear at low frequency and logarithmic at high frequency to suit the human hearing system.

The Mel scale is represented by the following equation

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where f is the frequency of the signal.

Linear prediction cepstral coefficients (LPCC) [10] use all the steps of LPC. The LPC coefficients are converted into cepstral coefficients using the following algorithm.

$$\mathrm{LPCC} = \mathrm{LPC}_i + \sum_{k=1}^{i-1}\left(\frac{k-i}{i}\right)\mathrm{LPCC}_{i-k}\mathrm{LPC}_k \tag{3}$$

PLP [11] uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: (1) the critical-bands spectral resolution, (2) the equal-loudness curve and (3) the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model. A fifth-order all-pole model is effective in suppressing speaker-dependent details of the auditory spectrum. In comparison with conventional LP analysis, PLP analysis is more consistent with human hearing.

The spectrum $P(\omega)$ of the original emotional speech signal is wrapped along its frequency axis ω into the Bark frequency Ω by

$$\Omega(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right] \tag{4}$$

where ω is the angular frequency in rad/s.

3.2 Emotion Classification

In this paper, multilayer perception (MLP) [12, 13] classifier is used and the results with different features are compared.

The structure of MLP for three layers is shown in Fig. 1. The three layers are input layer, hidden layer and output layer. Let each layer has its own index variable, '*k*' for output nodes, '*j*' for hidden nodes and '*i*' for input nodes. The input vector is propagated through a weight layer *V*. The output of *j*th hidden node is given by,

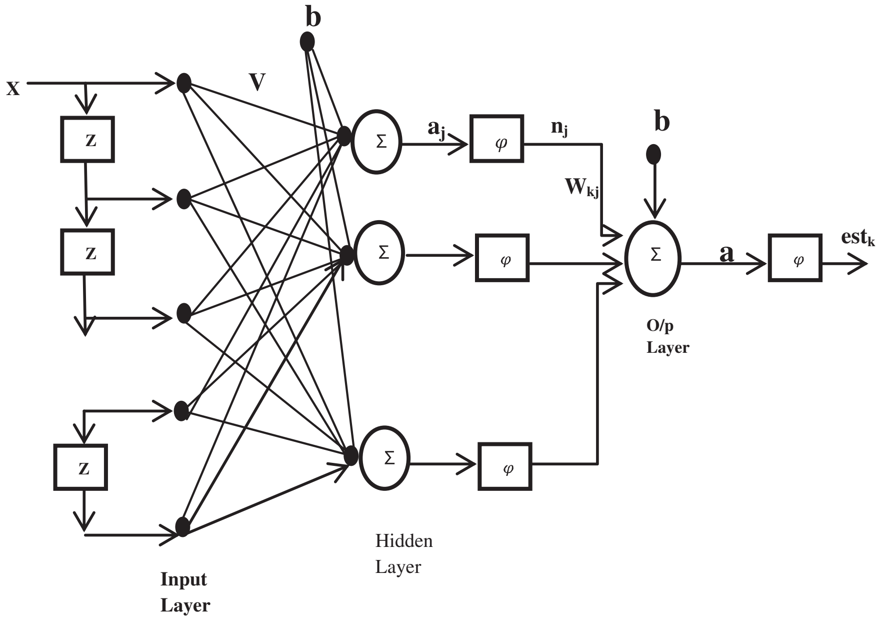

Fig. 1 Structure of MLP

$$n_j = \varphi\big(a_j(t)\big) \tag{5}$$

$$\text{where } a_j(t) = \sum_i x_i(t) v_{ji} + b_j \tag{6}$$

and a_j is output of jth hidden node before activation. x_i is the input value at ith node, b_j is the bias for jth hidden node, and φ is the activation function.

The output of the MLP network is determined by a set of output weights, W, and is computed as

$$\text{est}_k(t) = \varphi(a_k(t)) \tag{7}$$

$$a_k(t) = \sum_j n_j(t) w_{kj} + b_k \tag{8}$$

where η is the learning rate parameter of the back-propagation algorithm.

where est_k is the final estimated output of kth output node. The learning algorithm used in training the weights is back-propagation. In this algorithm, the correction to the synaptic weight is proportional to the negative gradient of the cost function with respect to that synaptic weight and is given as

$$\Delta W = -\eta \frac{\partial \xi}{\partial w} \tag{9}$$

where η is the learning rate parameter of the back-propagation algorithm.

4 Result and Discussion

The database has been prepared for four emotions for a group of 10 subjects for a sentence ‘*who is in the temple*’. The emotions include boredom, angry, sad and surprise. Database of children in the age group of six to thirteen were selected, including four boys and six girls. Duration of database is 1.5–4 s. The database was recorded using Nokia mobile and converted to wav file using format factory software at 8 kHz sampling frequency with 8 bits. From the Figs. 2, 3, 4 and 5, it was observed that high-arousal emotions like surprise and angry emotions have higher magnitudes than low-arousal emotions, like sad and boredom. Surprise emotions have highest magnitude, while bore emotions have lowest magnitude among all emotions.

As shown in the Table 1. The classification rate of MLP using MFCC feature vectors for the two classes of emotions was found to be highest (80 %) when all the four emotions angry, surprise, sad and bore are taken together. The reorganisation accuracy increases when one of the low-arousal emotions is compared with both the high-arousal emotions. Classification rate is lowest in case of LPC feature

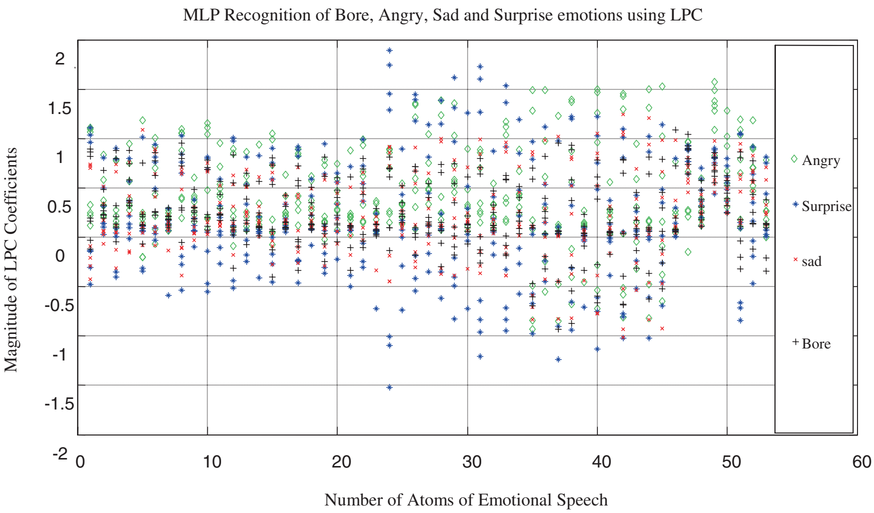

Fig. 2 Recognition of Bore, Angry, Sad and Surprise emotions using LPC

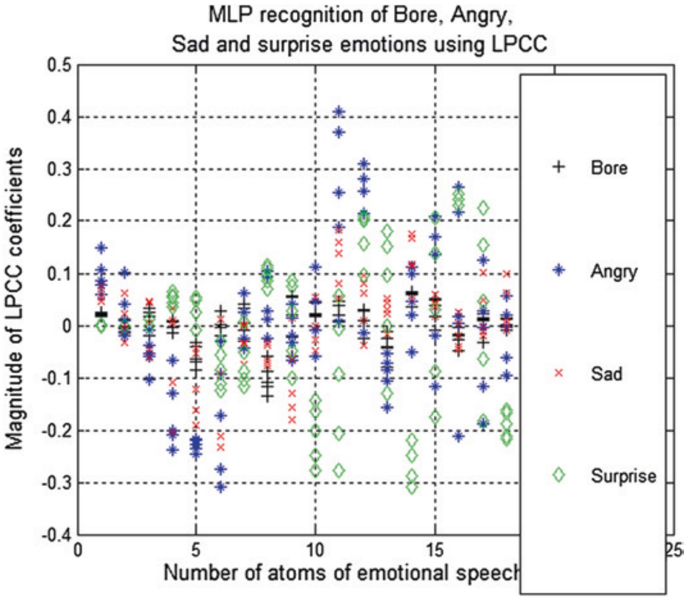

Fig. 3 Recognition of Bore, Angry, Sad and Surprise emotions using LPCC

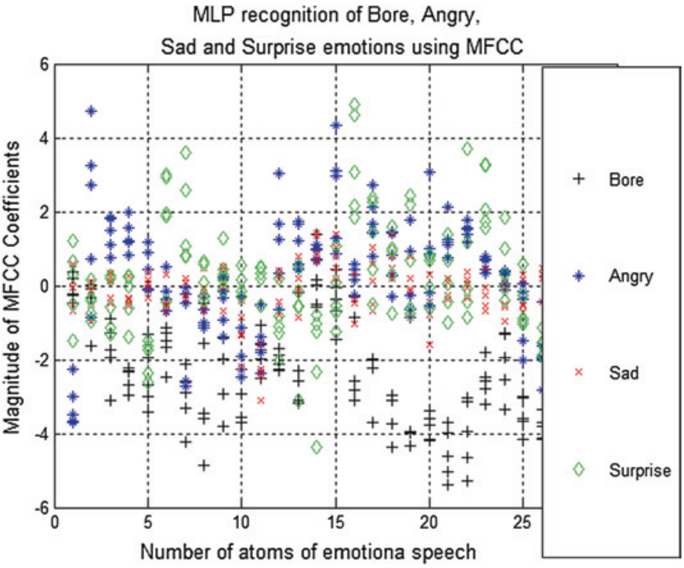

Fig. 4 Recognition of Bore, Angry, Sad and Surprise emotions using MFCC

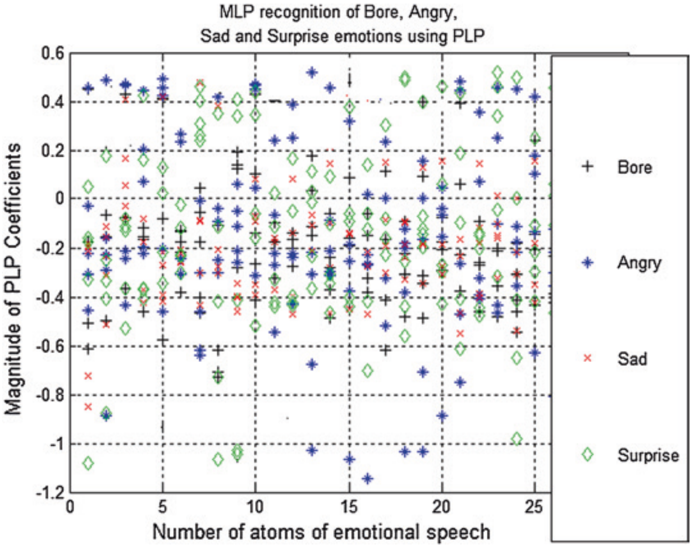

Fig. 5 Recognition of Bore, Angry, Sad and Surprise emotions using PLP

Table 1 Classification results

Feature extraction technique	Bore, angry, sad and surprise (%)	Bore, angry, surprise (%)	Bore, sad, surprise (%)
MFCC	80	83.20	81.40
LPC	48.60	56.20	52.70
LPCC	54.50	65.20	62.50
PLP	70.00	74.30	71.80

vectors, whereas LPCC gives better classification than LPC since it takes into account the cepstrum of the features. PLP features give better accuracy than both LPC and LPCC but gave poor performance than MFCC.

The superiority of MFCC and PLP is on account of their consideration of both linear and logarithmic scale of the voice range corresponding to the human hearing mechanism, while LPC and LPCC feature purely assume linear scale for the entire duration of speech.

5 Conclusion

It was observed that it is possible to distinguish the high-arousal speech emotions against the low-arousal emotion from their spatial representation as shown in Figs. 2, 3, 4 and 5. The range of magnitude of the feature coefficients can identify the emotions effectively as shown in these figures with maximum and minimum dispersions.

References

1. Mohanty, M., Mishra, A., Routray A.: A non-rigid motion estimation algorithm for yawn detection in human drivers. Int. J. Comput. Vision Robot. **1**(1), 89–109 (2009)
2. Mohanty, M.N., Routray, A., Kabisatpathy, P.: Voice detection using statistical method. Int. J. Engg. Techsci. **2**(1), 120–124 (2010)
3. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. **13**(2), (2005)
4. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: resources, features, and methods, speech communication. Elsevier **48**, 1162–1181 (2006)
5. Fragopanagos, N., Taylor, J.G.: Emotion recognition in human–computer interaction. Neural Networks, Elsevier **18**, 389–405 (2005)
6. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**, 561–580 (1975)
7. Ram, R., Palo, H.K., Mohanty, M.N.: Emotion recognition with speech for call centres using LPC and spectral analysis. Int. J. Adv. Comput. Res. **3**(3/11), 189–194 (2013)
8. Quatieri, T.F.: Discrete-Time Speech Signal Processing, 3rd edn. Prentice-Hall, New Jersey (1996)

9. Samal, A., Parida, D., Satpathy, M.R., Mohanty M.N.: On the use of MFCC feature vectors clustering for efficient text dependent speaker recognition. In: Proceedings of International Conference on Frontiers of Intelligent Computing: Theory and Application (FICTA)-2013, Advances in Intelligence System and Computing Series, vol. 247, pp. 305–312. Springer, Switzerland (2014)
10. Palo, H.K., Mohanty, M.N., Chandra M.: Design of neural network model for emotional speech recognition. In: International Conference on Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, April 2014
11. Hermansk, H.: Perceptual linear predictive (PLP) analysis of speech. J. Accoust. Soc. Am. **87**(4), 1739–1752 (1990)
12. Farrell, K.R., Mammone, R.J., Assaleh, K.T.: Speaker networks recognition using neural and conventional classifiers. IEEE Trans. Acoust. Speech Signal Process. **2**(1 part 11), (1994)
13. Javidi, M.M., Roshan, E.F.: Speech emotion recognition by using combinations of C5.0, neural network (NN), and support vector machines (SVM) classification methods. J. Math. Comput. Sci. **6**, 191–200 (2013)

Springer

http://www.springer.com/978-81-322-2195-1

Computational Vision and Robotics
Proceedings of ICCVR 2014
Sethi, I.K. (Ed.)
2015, XIII, 232 p. 110 illus., Softcover
ISBN: 978-81-322-2195-1