Improved Speech Emotion Classification from Spectral Coefficient Optimization

Inshirah Idris and Md Sah Salam

Abstract In order to improve the performance of speech emotion recognition systems, and to reduce the related computing complexity, this work proposed two approaches of spectral coefficient optimization. The two approaches are (1) optimized based on discrete spectral features and (1) combine spectral features. Experimental studies have been performed through the Berlin Emotional Database, using a support vector machine (SVM) classifier, and five spectral features including MFCC, LPC, LPCC, PLP and RASTA-PLP. The experiment results have shown that speech emotion recognition based on optimized coefficient numbers can effectively improve the performance. There were significant improvements in the accuracy 2 % for the first approach and 4 % for the second approach compared to that using the existing approaches. Moreover the second approach outperformed the first approach in the accuracy. This good accuracy came with reducing the features number.

Keywords Spectral features · Coefficients · MFCC · LPC · LPCC · PLP · RASTA-PLP · SVM

1 Introduction

Speech Emotion Recognition (SER) has become a hot research topic in recent years, due to its ability to identify the mood of a particular person from his or her voice. This makes it an important part of Human-Computer Interaction (HCI), as

I. Idris (✉)
Computer Science Department, Sudan University of Science and Technology, Khartoum, Sudan
e-mail: inshirah15@hotmail.com

M.S. Salam
Software Engineering Department, Universiti Teknologi Malaysia (UTM), Skudai, Johor, Malaysia
e-mail: sah@utm.my

P.J. Soh et al. (eds.), *Advances in Machine Learning and Signal Processing*, Lecture Notes in Electrical Engineering 387, DOI 10.1007/978-3-319-32213-1_22

used for many important applications including e-learning, robotics, healthcare, security, entertainment and so on. In general, SER is a pattern recognition system which uses a vector of extracted speech features from an emotional speech database, in order to recognize a persons emotional state, through the use of a classifier.

Since the feature extraction stage plays an important role in the performance of any pattern recognition system, the first issue in this area involves finding the best features that can help increase SER accuracy. Literature shows that there are four categories of acoustic speech emotion features, which include voice quality, prosodic, spectral and wavelet features. According to Wang et al. [1] the most commonly-used features include prosodic and spectral features.

When working with spectral features, including the Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC), the Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP), Relative Spectral Transform—Perceptual Linear Prediction (RASTA-PLP), the first and most important question is to determine how many coefficients are suitable for use. However there are no guidelines regarding how to choose the best number of coefficients. The tradeoffs in having large number of coefficients is that it may help to accommodate suitable features in the features vectors but it will also increase the feature dimensionality and possible redundancy which lead in increasing computational cost. On the other hand, small number of coefficients may lead to insufficient suitable features which may result in low recognition.

From the literature, researchers are used several number of coefficients in developing their SER systems (Fig. 1). Pierre-Yves [2] has used 10 MFCC coefficients. Rong et al. [3], Schuller et al. [4] and Lee et al. [5] have used 12 MFCC coefficients. Lee et al. [6], Wang and Guan [7] and Lugger and Yang [8] have used 13 MFCC coefficients. Schuller et al. [9] has used 15 MFCC coefficients. Several authors also have chosen to use the same number of coefficients for different

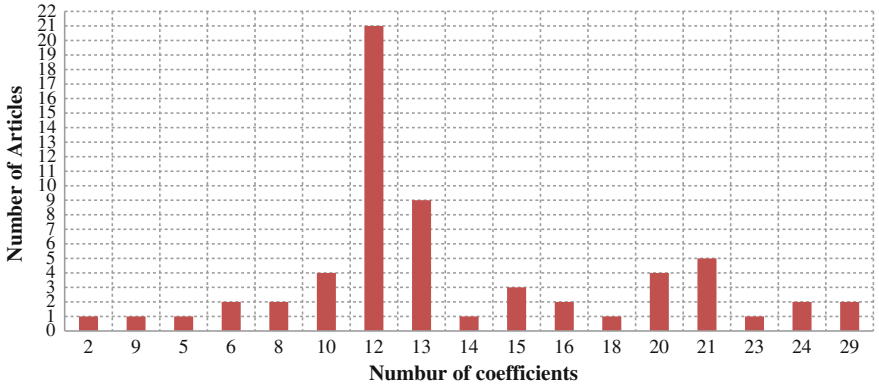

Fig. 1 Shows some coefficient numbers used by researchers in conjunction with spectral features for developing their systems, as obtained in a survey conducted between 2000 and 2015, with 40 papers

spectral features. For example, Kim et al. [10] has used 12 coefficients for both LPC and MFCC. Other researchers chose to use different numbers of coefficients for different spectral features. For example, Nwe et al. [11] chose to use 16 coefficients for LPCC, and 12 coefficients for both MFCC and LFPC. Fu et al. [12] also selected 10 coefficients for LPCC, and 12 coefficients for MFCC.

There are some researchers who also chose to test different numbers of coefficients for the same spectral features. For example, Koolagudi et al. [13] used 6, 8, 13, 21 and 29 coefficients for both LPCC and MFCC, while Murugappan et al. [14] used 13, 15 and 20 coefficients for MFCC, and Milton et al. [15] used MFCC with 10, 15, 24 and 23 coefficients. To reduce the dimensionality and computation of the SER system Hegde et al. [16] used the F-ratio technique to select a subset of 12 MFCC coefficients within the Hidden Markov Model (HMM), and concluded that the selection of 8 MFCC coefficients offers a better classification accuracy than that which could be achieved when selecting all 12 coefficients.

Based on the works mentioned above it is clear that there are no uniform patterns used to choose a suitable number of coefficients. This paper has proposed two approaches of selecting optimized numbers of coefficients, depending on the classifier, that could help to increase SER system accuracy while reducing feature vector dimensionality.

2 The Proposed System

Figure 2 shows the proposed speech emotion recognition system architecture, as based on optimized coefficients. The system process used is as follows:

1. The system starts with the speech records from the emotional database, which are described in Sect. 3.
2. The features step involves spectral features pre-processing and extracting using the selected scope number of coefficients, then the optimization of the number of coefficients for spectral features, the main method and algorithm described in Sect. 4.
3. After the optimizing process the features vectors are fed to the classifier, which provides the classification result (accuracy or class label). The classification method is described in Sect. 5.

3 The Berlin Emotional Database (EMO-DB)

A significant number of emotional speech databases have been developed for use when testing SER systems. Some of these databases are publicly available, while others have been created in order to meet a researchers particular needs. Emotional

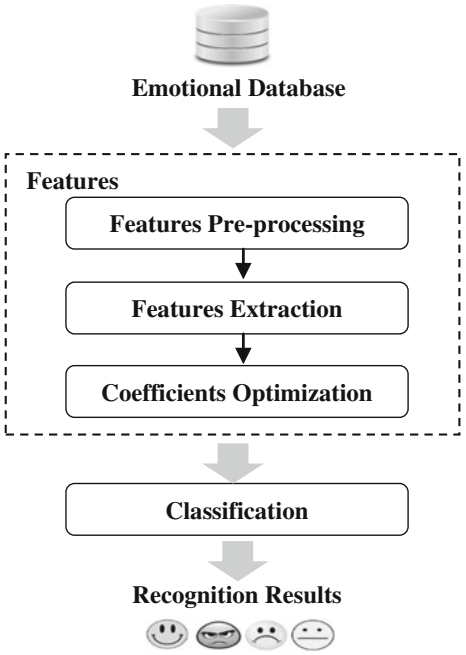

Fig. 2 The proposed system architecture

speech databases can be categorized into three different categories, namely acted, spontaneous and Wizard-of-Oz databases. It is more practical to use a database that has collected samples from real-life situations, and this can serve as a good baseline for creating real-life applications within a specific industry. However the acted database has been consider the easiest one to collect, and different studies have proven that it can offer strong results. It is therefore suitable for theoretical research.

Within this study, Berlin Emotional Database (EMO-DB) was selected as one of the most well-known acted emotional speech databases [17]. It also has been used with spectral features in many studies [18, 19]. The EMO-DB is an acted German emotional speech database recorded at the Department of Acoustic Technology, at TU-Berlin, and is funded by the German research community. It was recorded using a Sennheiser microphone set at a sampling frequency of 16 kHz, with the help of ten professional actors including five males and five females. These actors were asked to simulate seven emotions which included anger, boredom, disgust, fear, happiness, sadness and a neutral emotion, for ten utterances. Following the recording, twenty judges were asked to listen to the utterances in a random order in front of a computer monitor. They were allowed to listen to each sample only once, before they had to decide on the emotional state of the speaker. After the selection process, the database contained a total of 535 speech files.

4 Features

4.1 Features Pre-Processing and Extraction

In this work, we considered five different spectral features namely, MFCC, LPC, LPCC, PLP and RASTA-PLP. MFCC considered being the most used feature of speech [20–22]. It has been widely utilized within speech recognition and speech emotion recognition systems, and Poa et al. [23] reported it as the best and the most frequently acoustic features used in SER. LPC also has been considered one of the most dominant techniques for speech analysis [23]. LPCC is extension of the LPC that has the advantage of less computation, its algorithm is more efficient and it could describe the vowels in better manner [24].

PLP are also an improvement of LPC by using the perceptually based Bark filter bank. PLP analysis is computationally efficient and permits a compact representation [25]. While RASTA-PLP is improvement of the PLP method by adding a special band-pass filter was added to each frequency sub-band in traditional PLP algorithm in order to smooth out short-term noise variations and to remove any constant offset in the speech channel.

MATLAB R2012a was employed in order to compute 30 coefficients of the five features for a frame length of 25 ms every 10 ms, while ten different statistical measurements including minimum, maximum, stander deviation, median, mean, range, skewness, and kurtosis, were utilized for five spectral features from all speech samples.

4.2 Coefficients Optimization

Within this study, two approaches have been proposed for optimizing the number of coefficients for spectral features. The classifier has been used to compare a different number of coefficients, and then to select the coefficients that offer the best accuracy and the lowest number of features for speech emotion recognition. According to literature, the number of coefficients used in the past range from 2 to 29. From this the range of numbers of chosen coefficients was from 0 to 30 for MFCC, PLP and RASTA-PLP. However the first coefficients for LPC and LPCC have the same value for all records, namely 1 for LPC and −1 for LPCC, so the range of numbers of coefficients for both of them are chosen from 1 to 30. The coefficients optimization process as shown in Fig. 3 is as follows:

1. The first coefficient number in the search scope (0 for MFCC, PLP and RASTA-PLP and 1 for LPC and LPCC) has been chosen.
2. Then the features that corresponding to this coefficient number has been choosing from the extracted features vector.

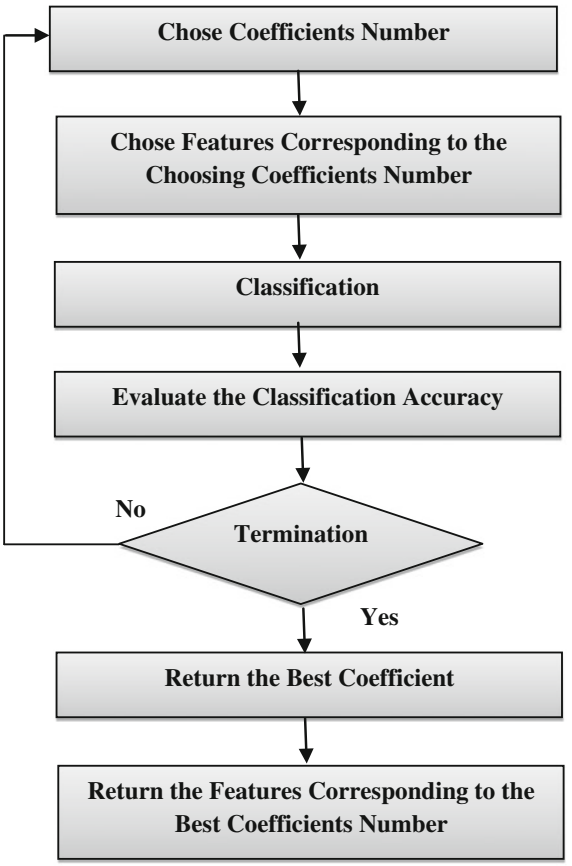

Fig. 3 The coefficients optimization process

3. Using SVM the accuracy of classification was calculated these steps are repeated until reaching the final number of coefficient number in the search scope (30 for the five features).
4. The coefficient numbers that give the highest accuracy with lowest number of features has been choosing, and the corresponding features have been choosing.
5. Finally the features have been combined in one vector.

The first approach was used to optimize the number of coefficients for the five features separately. The second approach was used to optimize the number of coefficients for the five features in a combination, The selection and evaluating of the coefficient number according to the classification accuracy have been done manually.

5 Classification

Several types of classifiers have been used in SER systems, including the Hidden Markov Model (HMM), the K-Nearest Neighbors (KNN), the Artificial Neural Network (ANN), the Gaussian Mixtures Model (GMM) and the Support Vector Machine (SVM). According to the literature [25] SVM and ANN are the most popular classifiers. Within this paper, SVM was adopted because it shows a strong performance when working with limited training data that has many features. SVM is a binary classifier used for classifications and regression. It can basically handle only two-class problems. SVM classifiers are mainly based on the use of kernel functions to nonlinearly map original features within a high dimensional space, in which data can be effectively classified using a linear classifier.

Classification with all speech utterances and spectral features was performed through the use of MATLAB R2012a. The radial basis kernel function (RBF) was employed with optimized g (in Gaussian function) and C (penalty parameter). The optimization of these classifier parameters was used in order to improve classifier accuracy. The scope of g is he scope of g is 2(−10:1:10) and the scope of C is 2 (−5:1:5). 5-fold. Cross-validation was performed for parameters selection. The performance analysis was undertaken using accuracy, which is the percentage of correctly-classified instances over the total number of instances.

6 Experiments and Analysis of Results

6.1 Optimized Based on Discrete Spectral Features

Within the first approach, the coefficients were separately optimized for the features, and the accuracy of the individual features was calculated. The result is shown in Fig. 4, where the x-axis indicates the number of coefficients, and the y-axis indicates the corresponding accuracy value. From the figure it can be observed that LPC gives the best accuracy of 58 % with 5 coefficients, and LPCC gives the best accuracy of 74 % with 12 coefficients.

For MFCC, as Fig. 5 shows, the best accuracy was 86 % with 20 coefficients. PLP gives the best results with 15 coefficients with an accuracy of 62 %, and finally RASTA-PLP gives the best accuracy of 54 % with 4 coefficients.

The results show that the MFCC feature provides the best accuracy among all features. This good result relates to the largest number of coefficients. LPCC and PLP provide good accuracy, with a reasonable number of coefficients. LPC and RASTA-PLP give the lowest numbers of coefficients and the worst accuracy. After separately determining the best coefficient values for every feature, the five features were combined. This provided an overall accuracy of 84 %, with 437 features.

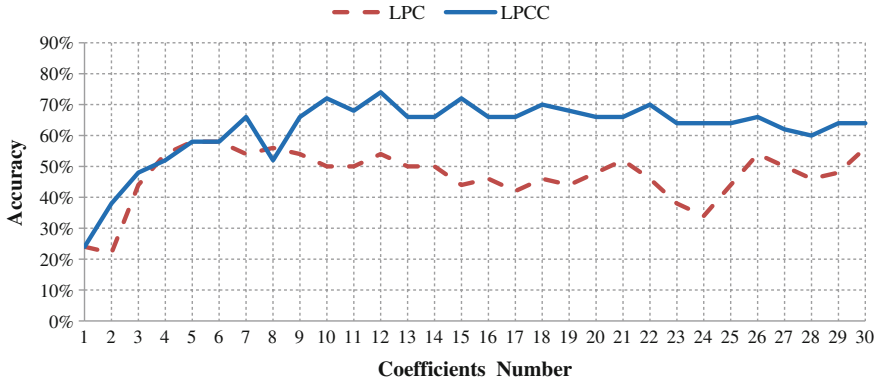

Fig. 4 The accuracy of LPC and LPCC for numbers of coefficients from 1 to 30

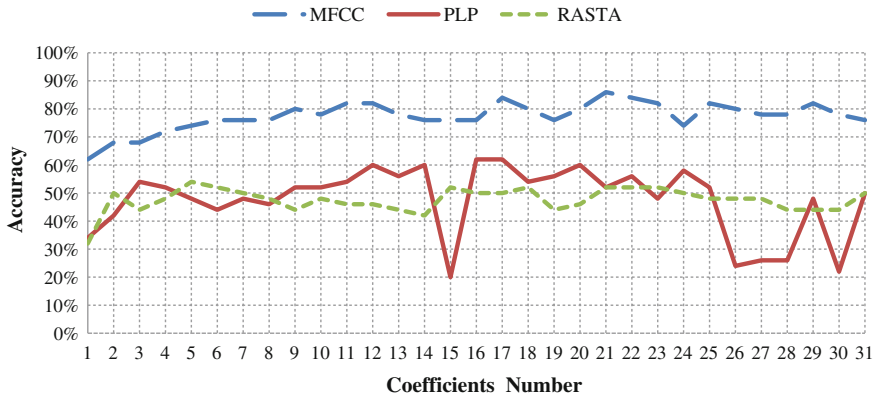

Fig. 5 The accuracy of MFCC, PLP and RASTA-PLP for all numbers of coefficients from 0 to 30

6.2 Optimized Based on Combine Spectral Features

Within the second model, the five features were combined first before coefficients optimization. Figure 6 showed that the best accuracy for the combined features was 88 % with 8 coefficients and 286 features.

The two approaches offered remarkable results as shown in Table 1. However, the second approach offered the highest accuracy with the lowest number of features.

When compared this study method undertaken with the greatest number of coefficients used in the past, namely 12 and 13 coefficients, as shown in Fig. 1, the result in Table 2 has shown that the number of coefficients selected by the two proposed Approaches can offer much greater accuracy than the number of coefficients used in the past. Additionally, the greater accuracy came with fewer features.

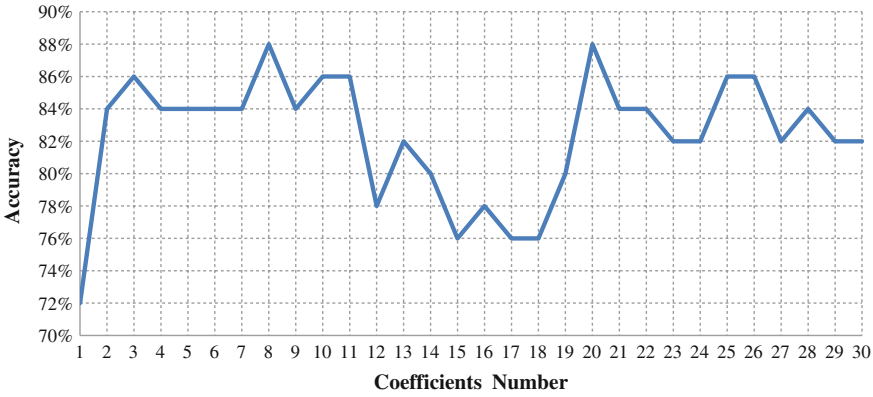

Fig. 6 The accuracy of the combined features for all numbers of coefficients, from 1 to 30

Table 1 Approaches accuracy

Approaches	Number of coefficients	Accuracy (%)	Number of features
Approach (1)	LPC(5), LPCC(12), MFCC(20), PLP(15), RASTA-PLP(4)	84	437
Approach (2)	8	88	286

Table 2 Comparison with the most number of coefficients used in SER

Number of coefficients	Accuracy (%)	Number of features
12	78	414
13	82	446

7 Conclusion

In this paper, two approaches for optimizing the coefficients numbers of spectral features, and for establishing a speech emotion model based on optimized coefficients, were proposed. Experiments have shown that the methods utilized for optimizing coefficients numbers not only increase the accuracy of the system when compared to the most commonly-used coefficients, but also reduces the numbers of features. This also shows that optimizing coefficient numbers for spectral features in combined, results in fewer features and better performance in speech emotion recognition, than when it is optimized separately before combination. Other Approaches used to optimize coefficients numbers will be studied in future works.

References

1. Wang F, Sahli H, Gao J, Jiang D, Verhelst W (2014) Relevance units machine based dimensional and continuous speech emotion prediction. Multimedia Tools Appl 1–18
2. Pierre-Yves O (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Stud 59(1):157–183
3. Rong J, Li G, Chen Y-PP (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45(3):315–328
4. Schuller B, Steidl S, Batliner A (2009) The inter-speech 2009 emotion challenge. In: INTERSPEECH, vol 2009. Citeseer, pp 312–315
5. Lee C-C, Mower E, Busso C, Lee S, Narayanan S (2011) Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 53(9):1162–1171
6. Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S (2004) Emotion recognition based on phoneme classes. In: INTER-SPEECH, pp 205–211
7. Wang Y, Guan L (2005) Recognizing human emotion from audiovisual information. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'05), vol 2. IEEE, pp ii–1125
8. Lugger M, Yang B (2008) Psychological motivated multi-stage emotion classification exploiting voice quality features. Speech Recognition, In-Tech, pp 395–410
9. Schuller B, Muller R, Lang MK, Rigoll G (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: INTERSPEECH, pp 805–808
10. Kim EH, Hyun KH, Kim SH, Kwak YK (2007) Speech emotion recognition using eigen-FFT in clean and noisy environments. In: RO-MAN 2007 the 16th IEEE international symposium on robot and human interactive communication. IEEE, pp 689–694
11. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden markov models. Speech Commun 41(4):603–623
12. Fu L, Mao X, Chen L (2008) Speaker independent emotion recognition based on SVM/HMMS fusion system. In: International conference on audio, language and image processing, ICALIP 2008. IEEE, pp 61–65
13. Koolagudi SG, Barthwal A, Devliyal S, Rao KS (2012) Real life emotion classification using spectral features and gaussian mixture models. Procedia Eng 38:3892–3899
14. Murugappan M, Baharuddin NQI, Jerritta S (2012) DWT and MFCC based human emotional speech classification using LDA. In: 2012 International conference on biomedical engineering (ICoBE). IEEE, pp 203–206
15. Milton A, Roy SS, Selvi S (2013) SVM scheme for speech emotion recognition using MFCC feature. Int J Comput Appl 69(9)
16. Hegde S, Achary K, Shetty S (2015) Feature selection using fisher's ratio technique for automatic speech recognition. arXiv preprint arXiv:1505.03239
17. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of german emotional speech. In: Interspeech, vol 5, pp 1517–1520
18. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53(5):768–785
19. Zhang Q, An N, Wang K, Ren F, Li L (2013) Speech emotion recognition using combination of features. In: 2013 fourth international conference on intelligent control and information processing (ICICIP). IEEE, pp 523–528
20. Kockmann M, Burget L et al (2011) Application of speaker-and language identification state-of-the-art techniques for emotion recognition. Speech Commun 53(9):1172–1185
21. Waghmare VB, Deshmukh RR, Shrishrimal PP, Janvale GB (2014) Emotion recognition system from artificial marathi speech using MFCC and LDA techniques. In: Fifth international conference on advances in communication, network, and computing, CNC, 2014

22. Kuchibhotla S, Vankayalapati H, Vaddi R, Anne K (2014) A comparative analysis of classifiers in emotion recognition through acoustic features. Int J Speech Technol 17 (4):401–408
23. Pao T-L, Chen Y-T, Yeh J-H, Liao W-Y (2005) Combining acoustic features for improved emotion recognition in mandarin speech. In: Affective computing and intelligent interaction. Springer, pp 279–285
24. Ingale AB, Chaudhari D (2012) Speech emotion recognition. Int J Soft Comput Eng (IJSCE) 2231–2307. ISSN
25. Pao T-L, Chen Y-T, Yeh J-H, Li P-J (2006) Mandarin emotional speech recognition based on SVM and NN. In: 18th international conference on pattern recognition ICPR 2006, vol 1. IEEE, pp 1096–1100