EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY Open Access

Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques

Tong Liu[1] and Xiaochen Yuan[1*]

Abstract

Emotion plays a dominant role in speech. The same utterance with different emotions can lead to a completely different meaning. The ability to perform various of emotion during speaking is also one of the typical characters of human. In this case, technology trends to develop advanced speech emotion classification algorithms in the demand of enhancing the interaction between computer and human beings. This paper proposes a speech emotion classification approach based on the paralinguistic and spectral features extraction. The Mel-frequency cepstral coefficients (MFCC) are extracted as spectral feature, and openSMILE is employed to extract the paralinguistic feature. The machine learning techniques multi-layer perceptron classifier and support vector machines are respectively applied into the extracted features for the classification of the speech emotions. We have conducted experiments on the Berlin database to evaluate the performance of the proposed approach. Experimental results show that the proposed approach achieves satisfied performances. Comparisons are conducted in clean condition and noisy condition respectively, and the results indicate better performance of the proposed scheme.

Keywords Speech emotion classification, Paralinguistic features, Spectral features, Support vector machine, Multi-layer perceptron classifier

1 Introduction

In recent years, the rising demand for artificial intelligence has made human-computer interaction program a heat aspect in computer technology. In this case, human emotion, specifically speech emotion detection, has raised great attention. The same utterance with different emotions can lead to a completely different meaning. The ability to perform various of emotion during speaking is also one of the typical characters of human. Therefore, technology trends to develop advanced speech emotion recognition systems in the demand of enhancing the interaction between computer and human beings, and thus emotion classification/recognition gradually become an undeniably essential application in voice signal processing and human-computer interaction [1–4].

Actually, the emotion classification is based on kinds of psychology models, which classify emotions according to certain rules and principles, helping us better understand and explain emotions. Here are several commonly used emotion classification models. The most basic one is the Six Basic Emotions Model [5], which categorizes emotions into six types: anger, disgust, fear, happiness, sadness, and surprise. The Six Basic Emotions Model (SBEM) is one of the most well-known and widely used models in emotion classification and has been applied in fields such as psychology, neuroscience, and computer graphics. However, this model is limited by oversimplification, as emotions are complex and some cannot be classified using this model. Therefore, a 2D emotion model, Circumplex Model of Affect [6], was proposed, where the emotion is presented by arousal and valence. Thereinto emotional

*Correspondence:
Xiaochen Yuan
xcyuan@mpu.edu.mo
[1] Faculty of Applied Sciences, Macao Polytechnic University, Macau, China

arousal refers to the intensity or degree of emotion, and emotional valence refers to the physiological response and behavioral tendency of emotion. Figure 1 displays the Circumplex Model of Affect applied on the Berlin Dataset emotion classes [7]. On this basis, there is another common emotion classification model called the three-factor emotion model [8]. In addition to emotional arousal and valence, the model adds emotional significance, which can better explain the performance and changes of emotions in different contexts.

In general, speech emotion classification systems consist of two parts: feature extraction and feature classification. Feature extraction deals with extracting speech features related to emotions using some extractor tools. In feature classification, it processes the extracted features to train the model using a classifier to predict the category of the emotion. In this way, the selection of emotion feature became the key point. Various speech features have been used in emotion classification [9, 10]. It has been reported there are three classes of traditional features based on the way of human to speak: prosodic features [11], syllable features [12], and spectral features [13, 14]. The prosodic feature is also called suprasegmental feature or paralinguistic feature [11]. It is mainly controlled by the pitch, energy, and frequencies of speech [15]. Syllable features are mostly associated with the quality of voice. Spectral feature is considered as the reflection of the relation between moving sound track and making sound. At most time, the linear prediction cepstrum coefficients (LPCC) [16] and the Mel-scale frequency cepstral coefficients (MFCC) [17] are on behalf of this genre of feature. Many researches have been done using various of features. Nwe et al. [18] proposed a speech emotion classification based on Markov model using short time log frequency power coefficient (LFPC). They proposed text independent method of emotion classification of speech, which made use of LFPC to represent the speech signals and used a discrete hidden Markov model (HMM) as the classifier. The performance obtained from their method had been compared with that of the LPCC [16] and MFCC feature commonly used in speech classification systems. Wong et al. [2] did a research about speech emotion classification by using a proposed Fourier parameter. This parameter is closely related to speech quality and used its perceptual content with the first and second-order differences for speaker-independent speech emotion classification. Due to the differences in speech characteristics between children and adults, with children's voices typically having higher pitch and faster speech rate, Albu et al. [19] utilized MFCC and some parameters to obtain the

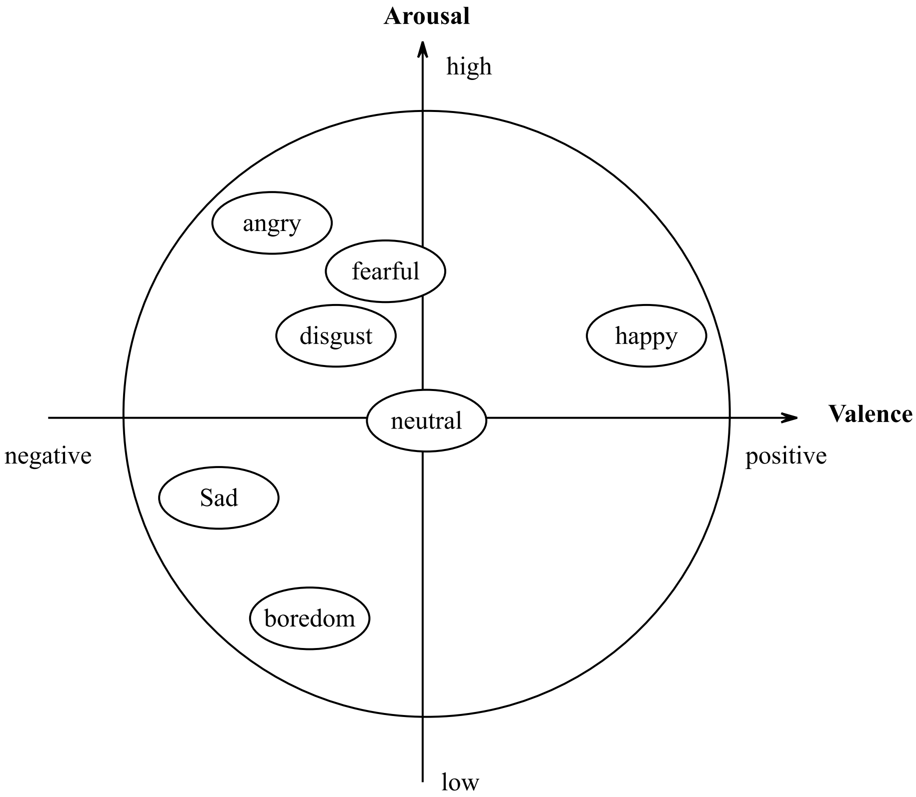


Fig. 1 Mapping the Berlin Dataset emotion classes onto the Circumplex Model of Affect [7]

feature vector for the neural network's input. These additional parameters include the short-term energy, the zero-crossing rate, the spectral roll-off, and spectral centroid. The mean and standard deviation of these features were then calculated and combined to create the final feature vector for speech signals.

In the aspect of classification method, various options can be taken for speech emotion system [20, 21]. Particularly, extreme learning machine (ELM) [20], K-nearest neighbor (KNN) [21], and support vector machine (SVM) [22] are all identical and widely used classifiers. Each classifier has its own properties that are suitable for a certain type of application. To enhance the performance of classification, a wide range of feature selection methods is optional. Deep spectral networks (DNNs) have also been used in classification of speech emotion. As presented in [23], an emotion state probability distribution for each speech segment was produced using DNNs, and utterance-level features were constructed from segment-level probability distributions. As for children's emotion classification, Albu et al. [19] tested the performance using radial basis function (RBF), ELM, and online sequential ELM (OS-ELM) networks. The experimental results indicate that the RBF network achieves superior performance compared to ELM.

However, though numerous researches have been done in years, the classification using these features under noisy environment is still a huge problem. The performance of classification will be teared down with noise-distorted signals. Therefore, in this study, a traditional paralinguistic feature mentioned as above and spectral feature are extracted to process classification procedure. The spectral feature is presented by a computational model which output a series of responses of a speech's particular characteristic frequency through auditory nerve fiber. The MFCC coefficients are extracted as spectral feature. Meanwhile, we propose to extract the traditional paralinguistic features presented by the INTERSPEECH 2013 paralinguistic challenge set [24] from the speech for speech emotion classification, using OpenSMILE [25] toolkit. A wide range of paralinguistic features and prosodic features are included in this set. The INTERSPEECH 2013 contains 6373 features, LLD including energy, spectrum, cepstrum, sound, log harmonic noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral clarity. To extract the features, the following procedures are conducted: extracting the low-level descriptors, adding Hamming windows, smoothing features, adding coefficients, and applying functionals to each of the descriptors.

After extracting the features, we employ the multi-layer perceptron classifier (MLP classifier) [26] and SVM respectively for the classification of speech emotions. The Berlin database is used to evaluate the performance, and the experimental results show that the proposed approach achieves good performances under different conditions and performs better than the related work in terms of the various evaluation metrics. The structure of this paper is organized as follows: Section 2 explains the proposed approach in detail, Section 3 illustrates the experimental results, and Section 4 concludes the paper and gives the future works.

2 Proposed speech emotion classification approach

It is important to mention that the test database has silent segments; therefore, the data is required to be pre-processed in order to make the speech samples suitable for the method in this study; meanwhile, pre-processing also has the ability to improve the accuracy of classification results. In this study, the sample rate of speech will be extracted by Librosa [27], which is a toolbox provided by Python, and mainly used in processing audio or analyzing audio. The silent segments will be removed by toolbox IPython, which uses the sample rate to specify the useful rate and remove the useless rate. Figure 2 shows the example of speech spectrum before and after pre-processing. (a) shows the spectrum of the original speech sample, and (b) shows the sample after pre-processing. After preprocessing, we employ the OpenSMILE and MFCC for feature extraction respectively. OpenSMILE is an open-source software for automatically extracting features from audio signals and classifying speech and music signals. The standard feature set used in this study is the openSMILE/openEAR "emobase" set, which has 998 acoustic features for emotion classification, including the different low-level descriptors (LLDs): intensity, loudness, 12 MFCC, pitch (F0), voiced probability, F0 envelope, 8 line spectral frequency, zero-crossing rate, and the delta regression coefficients.

MFCC is a feature widely used in automatic speech and speaker classification as they are suitable for understanding humans and the frequency at which humans speak. MFCC feature extraction consists of two key steps: Mel frequency analysis and cepstral analysis. MFCC is a set of key coefficients used to establish Mel cepstral. From the segments in the audio signal, we can obtain a set of cepstrums that are sufficient to represent the audio signal. Different from the general cepstrum, the frequency band on the Mel cepstrum is evenly distributed on the Mel scale, such a frequency band will be more linear than what we generally see. The cepstral representation method is closer to the human nonlinear auditory system. Therefore, in this method, we choose MFCC as the feature extraction method.

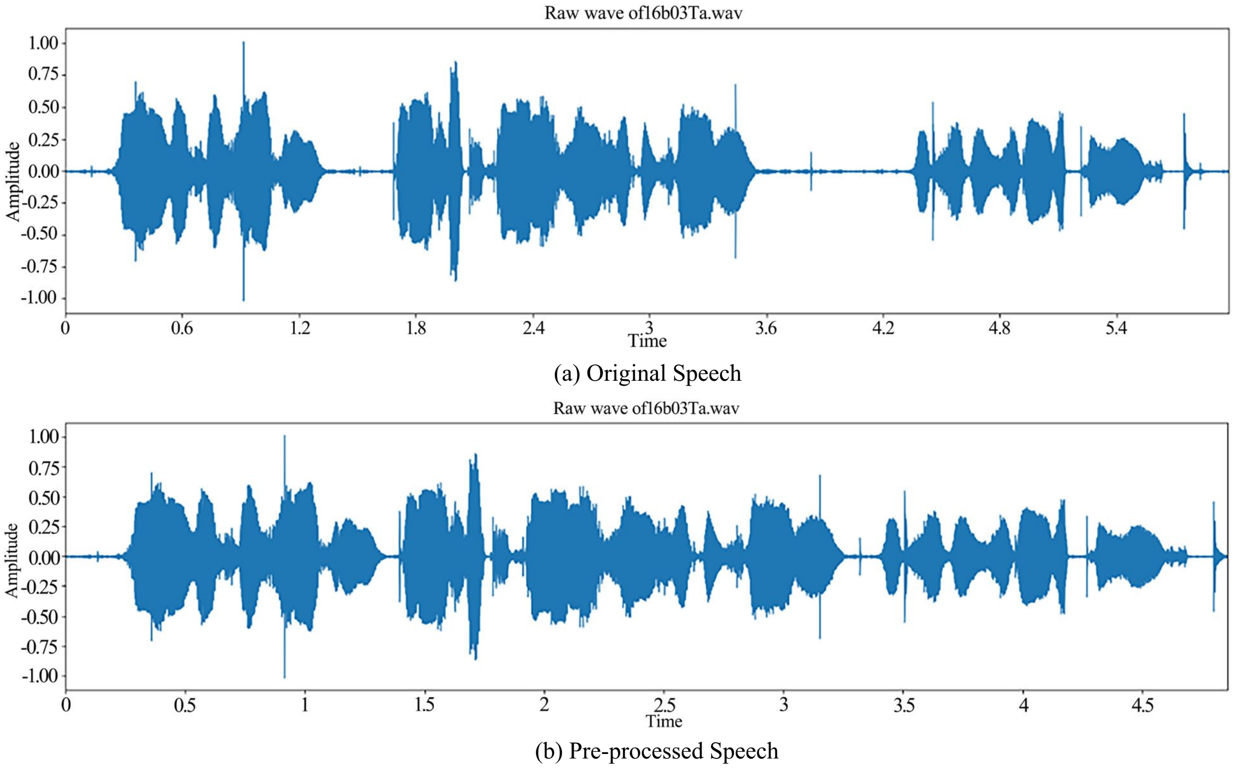


Fig. 2 Example of speech spectrum after pre-processing

2.1 Classification using MLP classifier

Figure 3 shows the flowchart of the proposed SEC-MLP. The MLP classifier is a network made up of perceptron. The structure consists of an input layer L_{in}, n hidden layers L_h, and an output layer L_{out}. Each layer consists of several neurons, and each neuron in a layer is fully connected to every neuron in the adjacent layers [28]. The L_{in} is the first layer of a neural network, which receives input data and passes it to the next layer. And each neuron corresponds to one feature of the inputs; therefore, the number of neurons N_{in} in L_{in} depends on the number of input features. The output layer L_{out} is used for making predictions for the given input. Each neuron in L_{out} represents a class and outputs the probability of belonging to its corresponding class. Generally, the number of neurons in L_{out} equals the number of classes in a classification problem. If the input needs to be classified into N_{out} classes, then the output layer should have N_{out} neurons. The layers present in between the input layer and output layer are called hidden layers L_h. Each L_h of a neural network receives signals from the previous layer's outputs and passes them on to the next layer as inputs through weighted connections. Each L_h also has N_h neurons, and the number of n and N_h can be adjusted according to task requirements. For speech emotion recognition tasks, the input features are typical with high dimensionality, while the number of output classes is relatively small. Therefore, selecting a smaller number of neurons can help reduce the neural network's complexity and decrease the training time and memory usage of the model [29, 30]. In this paper, we set the number of neurons in each L_h to 300. Given k training samples, an MLP with an input feature dimension of N_{in}, n hidden layers with N_h neurons in each hidden layer, and N_{out} output neurons, the time complexity for classification in one iteration is $O(k \times N_{in} \times {N_h}^n \times N_{out})$.

In addition, data augmentation is employed before applying the MLP classifier. Data augmentation is a set of techniques to artificially increase the amount of data by generating new data points from existing data. This includes making small changes to data or using deep learning models to generate new data points. It is useful to improve performance and outcomes of machine learning models by forming new and different examples to train datasets. There are many different methods to achieve data augmentation, such as padding, cropping, random erasing, adding white noise and pitch tuning, and so on. In this work, we use five data augmentations, including adding white noise, pitch tuning, random shifting, peed and pitch tuning, and stretching

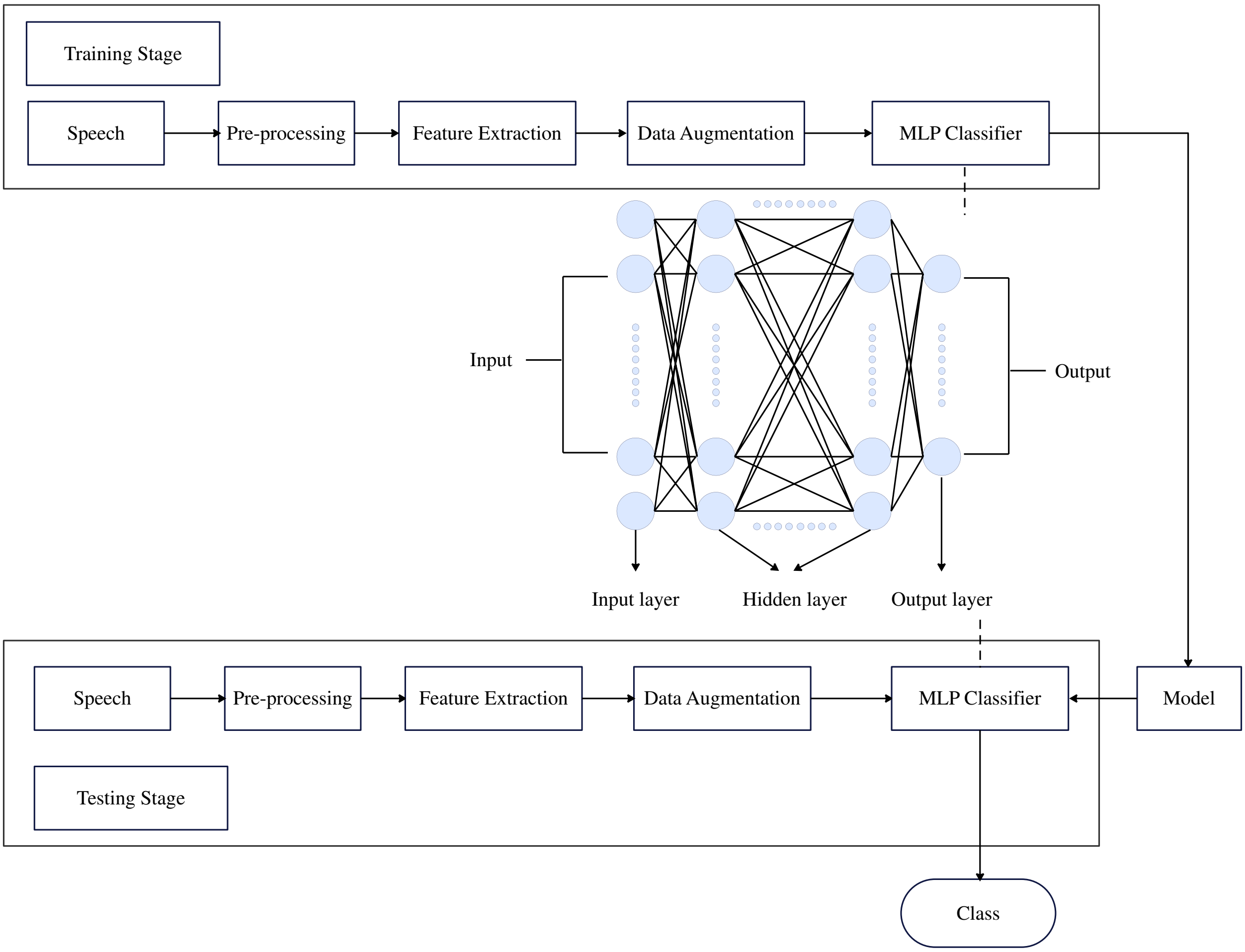


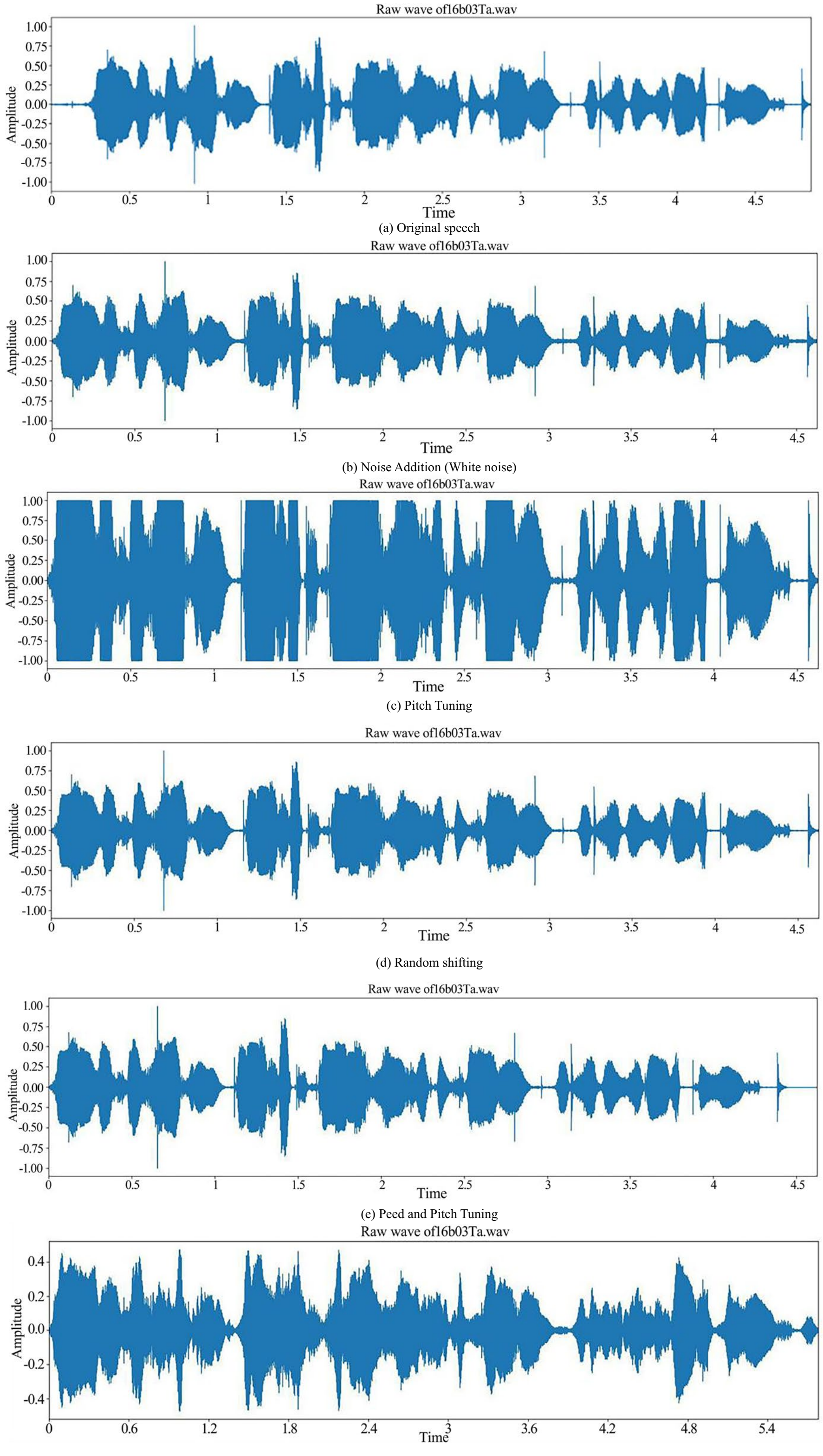
Fig. 3 Flow chart of speech emotion classification using MLP classifier (SEC-MLP)

the sound. Figure 4 shows the speech example and the spectrum after each data augmentation.

In the proposed SEC-MLP, we employ the fully connected spectral network MLP classifier for classification. MLP classifier relies on an underlying spectral network to perform the classification task. The MLP classifier is made to train on the given dataset. The training phase enables the MLP classifier to learn the correlation between the set of inputs and outputs. During training, the MLP classifier adjusts model parameters such as weights and biases in order to minimize the error. The MLP classifier uses backpropagation to make weight and bias adjustments relative to the error. During the implementation phase, the batch size is set as 256, and the Adam optimizer is adopted along with an adaptive learning rate approach. The learning rate is initialized to 0.001. If the learning rate cannot be decreased for two consecutive epochs, the current learning rate is divided by 5.

2.2 Classification using SVM

Figure 5 shows the flow chart of the proposed SEC-SVM. With the extracted feature as explained in Section 2.1, we apply the feature selection algorithm to enhance the performance. To select the features, various of subsets are generated from all features, and each subset will be filtered with learning algorithm and evaluated depending on its performance. It is the process of selecting highly relevant subset within a variety of features. The goal of feature selection algorithm is to find out the more relevant data so that the performance can be improved by deleting the low relevant features. With the feature selection, the data visualization and data understanding can be promoted, the measurement and storage requirement can be reduced, the training, testing, and utilization times can be reduced, the efficiency of operation can be increased, and ultimately the performance of classification can be improved. In this work, the performance of each algorithm is evaluated by the result of averaged



(a) Original speech

(b) Noise Addition (White noise)

(c) Pitch Tuning

(d) Random shifting

(e) Peed and Pitch Tuning

(f) Stretching

Fig. 4 Speech spectrum after data augmentation. **a** Original speech, **b** noise addition, **c** pitch tuning, **d** random shifting, **e** peed and pitch tuning, and **f** stretching

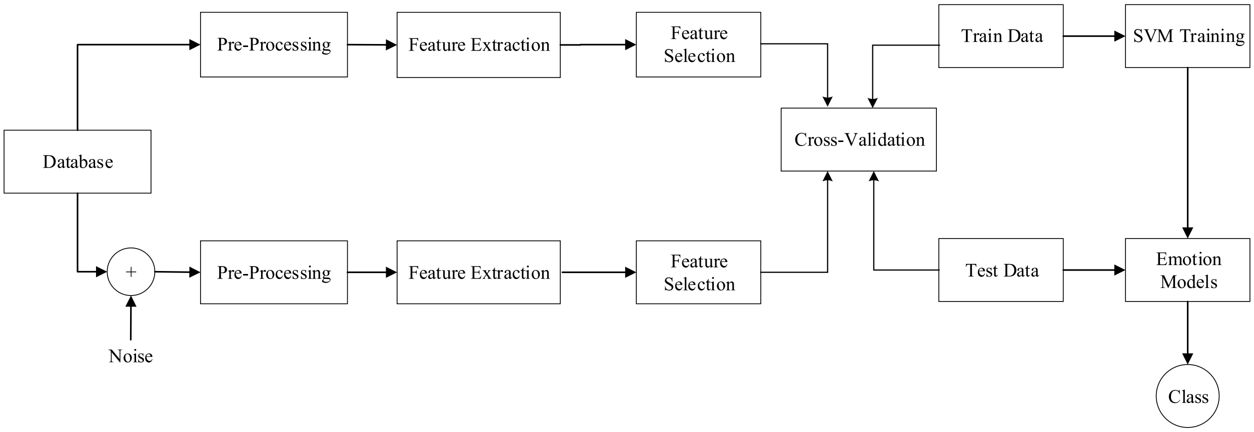


Fig. 5 Flow chart of speech emotion classification using SVM (SEC-SVM)

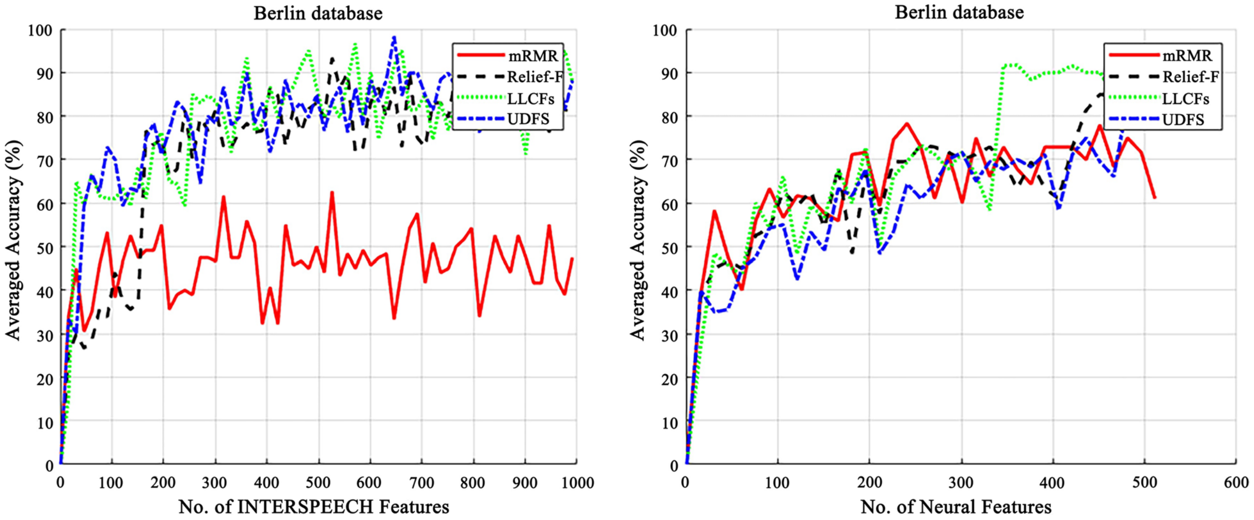


Fig. 6 Performance of different feature selection algorithms

accuracy using the selected feature selection method. Two categories of selection algorithms are used for evaluation: supervised algorithm and unsupervised algorithm [31]. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs [32]. Unsupervised learning is to model the hidden patterns or underlying structure in the given input data in order to learn about the data. The supervised algorithms are the minimum redundancy maximum relevance feature selection (mRMR) [33] and the Relief-F [34]. The unsupervised algorithms are the local learning-based clustering feature selection (LLCFs) [35] and unsupervised $L_{2,1}$-norm regularized discriminative feature selection (UDFS) [36]. We calculate the averaged ACC of the four feature selection methods respectively and show the results in Fig. 6, where the horizontal axis indicates the number of INTERSPEECH features extracted by OpenSMILE. It can be seen that the Relief-F, LLCFs LLCFs, and UDFS can achieve the better performance than mRMR.

In the SEC-SVM, the SVM classifier is applied for speech emotion classification. The SVM-train algorithm builds a model that assigns test examples to one category or another. First, the training labels, training characteristics, testing labels, and testing characteristics are generated. Next, an SVM model is built by using the above training labels and characteristics. The best constant and gamma coefficient for building training model is computed by grid search. Finally, SVM-predict defines the way of predicting labels and the predicting accuracy. Cross-validation algorithm is used in the SEC-SVM. The main idea of cross-validation is to divide original dataset

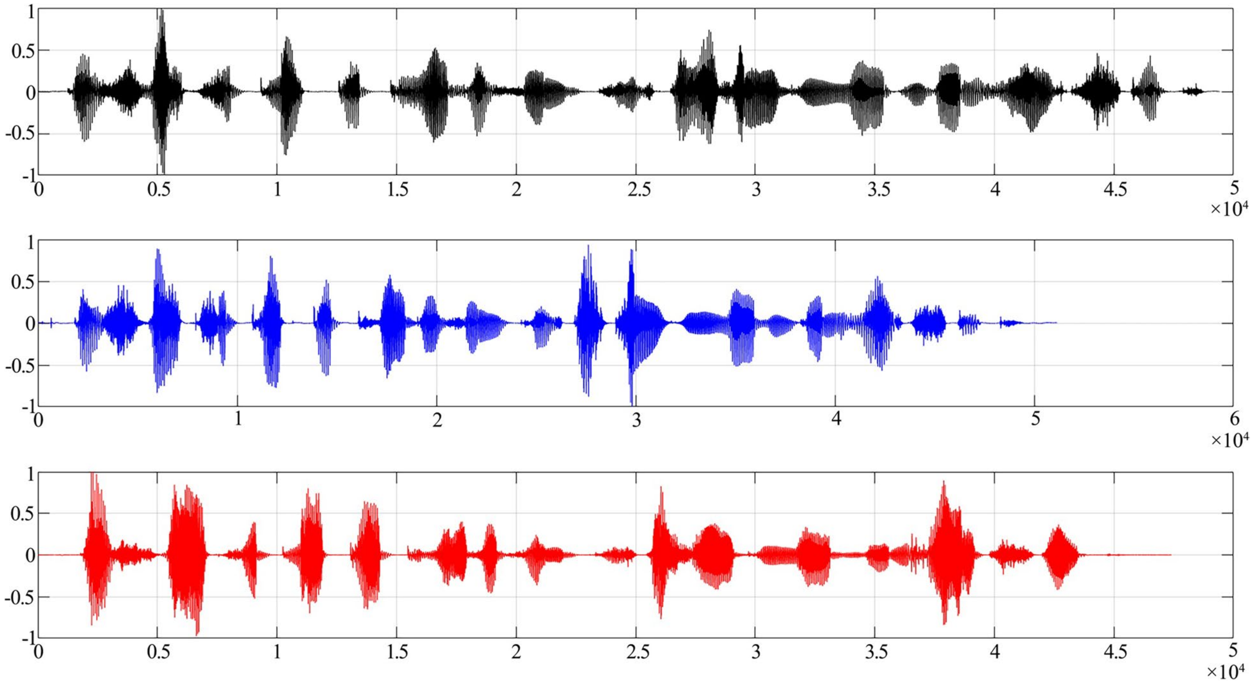


Fig. 7 The spectrum examples taken from Berlin database

into several parts; some of them is sent to classification method as training data and the other representing testing data. This procession has the ability to evaluate the performance of classification and flags problems like overfitting. In this study, the tenfold cross-validation, which is a non-exhaustive cross-validation method, is used to divide dataset into training and testing parts. The tenfold cross-validation separates dataset into ten approximately equal subsets. Nine of the subsets are set as the train data and the remaining one is set as the test data. The procession will repeat ten times to make sure each part of the ten subsets has been used as train data at least once. The distributed data is then sent for classification use. The whole procedure will be repeated 20 times. Given k training samples with m features, the time complexity for SVM is $O(k \times m^2)$.

Figure 6 contains four graphs, each of it shows the averaged accuracy got by using the four mentioned feature selections method. The result is obtained by filtering the whole feature set by four feature selection algorithms and then the filtered feature was sent to the SVM for classification. The x-axis of the graph represents the number of features, and the y-axis is the average accuracy. The goal of this test is to find out the best feature selection algorithm and the number of filtered features that can get the highest averaged classification accuracy. As presented in the figure, the best performance of averaged accuracy for the spectral feature is achieved by UDFS algorithm and the best performance of averaged accuracy for INTERSPEECH feature is achieved by LLCFs algorithm. By using UDFS algorithm for spectral feature and LLCFs algorithm for INTERSPEECH feature, the best features were selected and sent to SVM for classification.

3 Experimental results and discussions

In this study, the well-known database Berlin [37] was used to test the performance of the proposed method. The Berlin database [37] is a German database of emotional speech, which provides 535 audios of daily utterances containing seven emotions: (1) "anger," 2) "boredom," (3) "disgust," (4) "fearful," (5) "happy," (6) "neutral," and (7) "sad." The data was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz. Examples of Berlin database are presented in Fig. 7. The figure represents the same utterance spoken by a Germany in three different emotions. The spectrums on the top represent fear, the middle spectrum shows happiness, and the last is acted in anger.

To evaluate the performance of the proposed approach, the various metrics are calculated using (1) ~ (5), where accuracy (ACC) represents accuracy of the model [38,

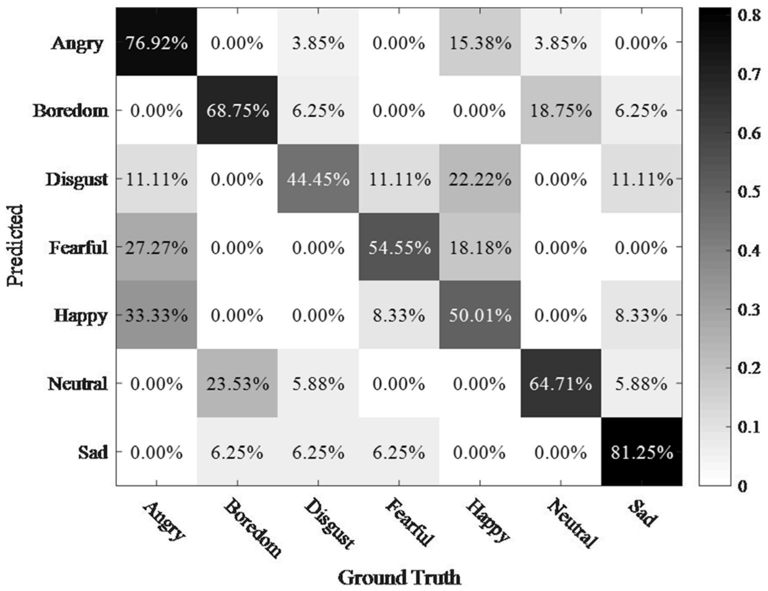


Fig. 8 Performance of SEC-MLP without data augmentation using paralinguistic feature. ACC = 62.95%, TPR = 39.72%, TNR = 79.26%, PPV = 57.36%, NPV = 65.18%. ACC results are used to construct the confusion matrix

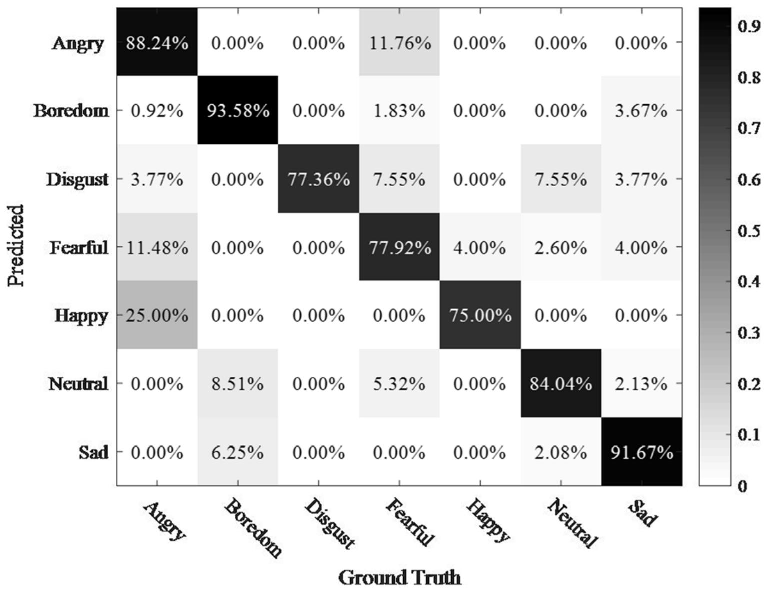


Fig. 9 Performance of SEC-MLP with data augmentation using paralinguistic feature. ACC = 83.97%, TPR = 69.07%, TNR = 91.28%, PPV = 79.52%, NPV = 85.75%. ACC results are used to construct the confusion matrix

39], true positive rate (TPR) measures how good the model is detecting positive events, and true negative rate (TNR) represents how apt the assignment to the positive class is. Positive prediction value (PPV) shows how exact the model is at assigning positive events to the positive class, and negative predictive value (NPV) measures how accurate the model is in detecting negative events. TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative predicted labels, respectively.

$$ACC = \frac{TP + TN}{P + N} \tag{1}$$

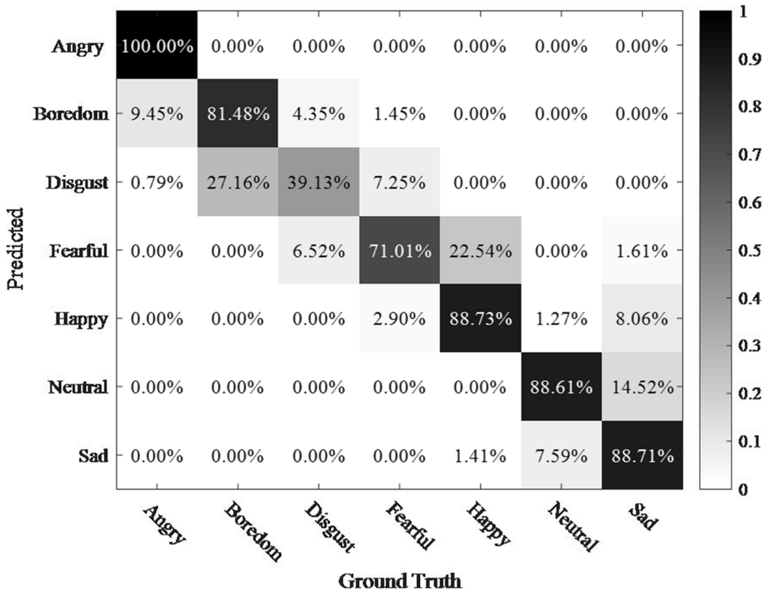


Fig. 10 Performance of SEC-SVM using paralinguistic feature. ACC = 83.74%, TPR = 90.71%, TNR = 96.13%, PPV = 99.03%, NPV = 96.11%. ACC results are used to construct the confusion matrix

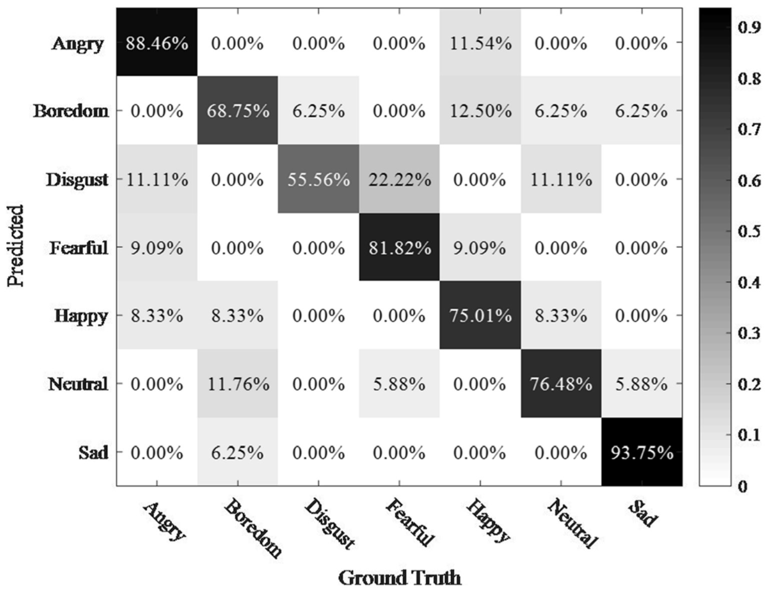


Fig. 11 Performance of SEC-MLP without data augmentation using spectral feature. ACC = 77.12%, TPR = 57.57%, TNR = 88.90%, PPV = 75.75%, NPV = 77.67%. ACC results are used to construct the confusion matrix

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2}$$

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

$$NPV = \frac{TN}{TN + FN} \tag{5}$$

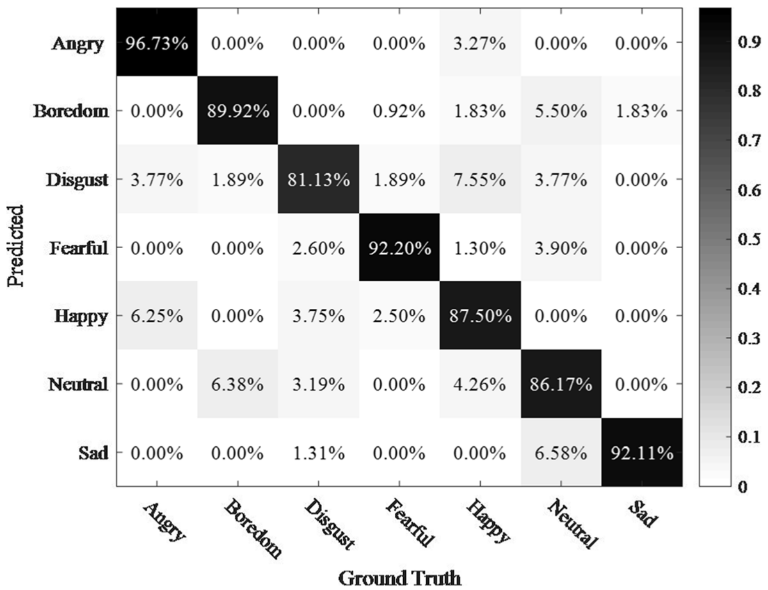


Fig. 12 Performance of SEC-MLP with data augmentation using spectral feature. ACC = 89.39%, TPR = 78.38%, TNR = 94.5%, PPV = 86.84%, NPV = 90.42%. ACC results are used to construct the confusion matrix

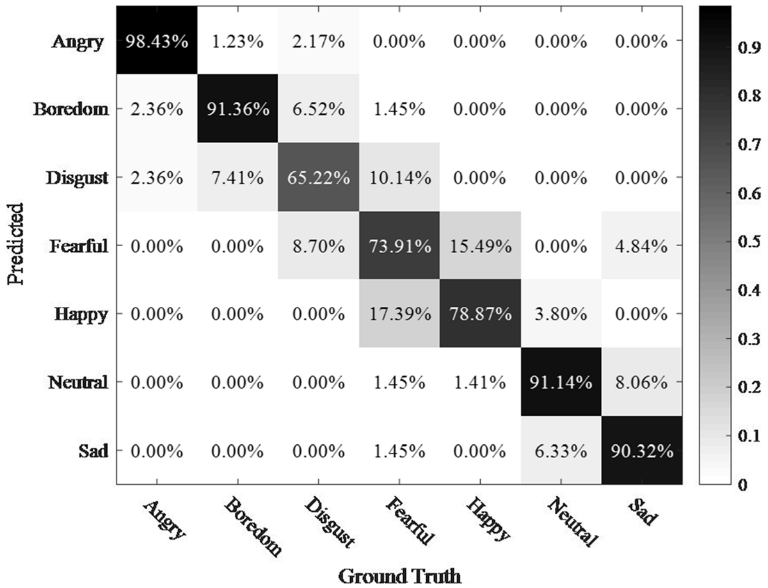


Fig. 13 Performance of SEC-SVM using spectral feature. ACC = 86.73%, TPR = 95.42%, TNR = 99.41%, PPV = 98.43%, NPV = 98.26%. ACC results are used to construct the confusion matrix

3.1 Performance of proposed approach using paralinguistic feature

We show the performance of proposed speech emotion classification approach using paralinguistic feature in this section, as shown in Figs. 8, 9, and 10, where the ACC results are calculated and are used to construct the confusion matrix. Figures 8 and 9 show the results of SEC-MLP without and with data augmentation respectively. The results indicate that performance of the approach can improve a lot with operations of data augmentation. In addition to the ACC results, we also calculate the TPR, TNR, PPV, and NPV results of each speech emotion classification, and the averaged results are calculated in terms of without data augmentation

Table 1 Comparison of proposed scheme with existing work under clean condition

		ACC	TPR	TNR	PPV	NPV
Method of [39]		78.07%	88.19%	96.60%	84.35%	97.64%
SEC-MLP using paralinguistic feature	Without data augmentation	62.95%	39.72%	79.26%	57.36%	65.18%
	With data augmentation	83.97%	69.07%	91.28%	79.52%	85.75%
SEC-SVM using paralinguistic feature		83.74%	90.71%	96.13%	**99.03%**	96.11%
SEC-MLP using spectral feature	Without data augmentation	77.12%	57.57%	88.90%	75.75%	77.67%
	With data augmentation	**89.39%**	78.38%	94.50%	86.84%	90.42%
SEC-SVM using spectral feature		86.73%	**95.42**%	**99.41**%	98.43%	**98.26%**

The bolded data indicates the best results

Table 2 Comparison of proposed scheme with existing work under noisy condition

SNR	Method	ACC	TPR	TNR	PPV	NPV
5dB	Method of [39]	**82.55%**	91.68%	93.72%	82.60%	97.36%
	SEC-SVM using paralinguistic feature	66.17%	**97.87%**	88.22%	72.44%	**99.24%**
	SEC-SVM using spectral feature	57.38%	69.40%	**100.00%**	**100.00%**	76.27%
10dB	Method of [39]	**84.52%**	93.29%	94.52%	84.62%	97.89%
	SEC-SVM using paralinguistic feature	67.48%	**99.04%**	91.49%	81.10%	**99.37**%
	SEC-SVM using spectral feature	71.59%	82.47%	**100.00%**	**100.00%**	90.46%
15dB	Method of [39]	**84.68%**	92.58%	94.62%	84.94%	**97.68%**
	SEC-SVM using paralinguistic feature	74.39%	**94.40%**	96.89%	92.91%	97.56%
	SEC-SVM using spectral feature	83.55%	92.70%	**99.00%**	**97.27%**	96.97%

The bolded data indicates the best results

and with data augmentation respectively. When without data augmentation, the averaged ACC, TPR, TNR, PPV, and NPV are 62.95%, 39.72%, 79.26%, 57.36%, and 65.18% respectively, while when data augmentation is applied, the averaged ACC, TPR, TNR, PPV, and NPV are 83.97%, 69.07%, 91.28%, 79.52%, and 85.75%. respectively. The overall results indicate that data augmentation greatly improves the classification results.

Figure 10 shows the performance of the proposed SEC-SVM, where the concatenated data from Berlin database was taken apart into ten subsets by tenfold cross-validation algorithm, and nine of the ten were transmitted as the input of SVM for train model. Predict label is generated by SVM predict after inserting test data, model, and predicting option. The averaged ACC, TPR, TNR, PPV, and NPV are calculated as 83.74%, 90.71%, 96.13%, 99.03%, and 96.11% respectively. It can be easily seen that the SEC-SVM performs better than the SEC-MLP in the same circumstances.

3.2 Performance of proposed approach using spectral feature

Similarly as in Section 3.1, in this section, we show the performance of proposed speech emotion classification approach using spectral feature, and the results are shown in Figs. 11, 12 and 13, where the ACC results are used to construct the confusion matrix. Figures 11 and 12 show the results of SEC-MLP without and with data augmentation respectively. We calculate the TPR, TNR, PPV, and NPV results of each speech emotion classification as well, and the averaged results are calculated in terms of without data augmentation and with data augmentation respectively. When without data augmentation, the averaged ACC, TPR, TNR, PPV, and NPV are 77.12%, 57.57%, 88.90%, 75.75%, and 77.67% respectively, while when data augmentation is applied, the averaged ACC, TPR, TNR, PPV, and NPV are 89.39%, 78.38%, 94.5%, 86.84%, and 90.42%. respectively. The results indicate that the approach performance can improve a lot with operations of data augmentation.

Figure 13 shows the performance of the proposed SEC-SVM, where the concatenated data from Berlin database was taken apart into ten subsets by tenfold cross-validation algorithm as well, and nine of the ten were transmitted as the input of SVM for train model. The averaged ACC, TPR, TNR, PPV, and NPV are calculated as 86.73%, 95.42%, 99.41%, 98.43%, and 98.26% respectively. It can be easily seen that the SEC-SVM performs better than the SEC-MLP in the same circumstances.

3.3 Comparison of the proposed scheme with the existing methods

To test the performance of proposed scheme under real speech environment, in this section, we compare performance of the proposed scheme with the existing method [39], when under clean condition and noisy conditions, respectively, and show the results in Tables 1 and 2. In Table 2, Gaussian white noise with different signal-to-noise ratios (5, 10, 15 dB) is used to distort testing speech samples before feature extraction. The ACC, TPR, TNR, PPV, and NPV are calculated to show the performance of the proposed scheme.

In Table 1, the performance of the proposed SEC-MLP and SEC-SVM are calculated, using paralinguistic feature and spectral feature, respectively. The metrics ACC, TPR, TNR, PPV, NPV are calculated, and the best result of each metric is highlighted in bold for clarity. The results indicate that SEC-MLP using paralinguistic feature achieves much better results with data augmentation, in terms of all ACC, TPR, TNR, PPV, and NPV. On the other hand, the SEC-SVM using spectral feature achieves better results in terms of TPR, TNR, and NPV. In Table 2, we simulate the real environment by adding Gaussian white noise with different signal-to-noise ratios (5, 10, 15 dB) and calculate the metrics ACC, TPR, TNR, PPV, and NPV respectively.

4 Conclusions and future works

In this paper, we propose the approach for speech emotion classification and we evaluate the approach in various cases. The traditional paralinguistic features based on the INTERSPEECH 2013 paralinguistic challenge set were extracted using OpenSMILE toolkit. The spectral features are extracted by the well-known MFCC. SVM and MLP classifier are employed respectively for the classification. Experiments have been conducted on the Berlin database to evaluate the performance of the proposed approach. The ACC, TPR, TNR, PPV, and NPV are respectively calculated to measure the proposed approach. Experimental results show that the proposed approach achieves good performances under different conditions and performs better than the related work in terms of the various evaluation metrics. Our future work will focus on combining the different features for better classification performance and applying deep learning techniques for classification.

Abbreviations

MFCC	Mel-frequency cepstral coefficients
SVM	Support vector machine
LSPC	Log frequency power coefficient
ELM	Extreme learning machine
MLP	Multi-layer perceptron
LLD	Low-level descriptors
SEC-MLP	Speech emotion classification using MLP classifier
SEC-SVM	Speech emotion classification using SVM
LLCF	Clustering feature selection
UDFS	Unsupervised discriminative feature selection
ACC	Accuracy
TPR	True positive rate
TNR	True negative rate
PPV	Positive prediction value
NPV	Negative predictive value

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization: Tong Liu and Xiaochen Yuan; data curation: Tong Liu; formal analysis: Tong Liu; funding acquisition: Xiaochen Yuan; methodology: Tong Liu and Xiaochen Yuan; project administration: Xiaochen Yuan; supervision: Xiaochen Yuan; validation: Tong Liu; visualization: Tong Liu; writing——original draft: Tong Liu; writing——review and editing: Xiaochen Yuan. The authors read and approved the final manuscript.

Funding
This research was funded by the Research project of the Macao Polytechnic University (Project No. RP/ESCA-03/2021, RP/FCA-12/2022).

Availability of data and materials
The dataset we used in this work is Berlin dataset, which is also known as EMODB database, and is available at: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 February 2023 Accepted: 29 April 2023
Published online: 15 May 2023

References

1. X. Cao, M. Jia, J. Ru, T.w. Pai, Cross-corpus speech emotion recognition using subspace learning and domain adaption. EURASIP J. Audio Speech Music Process. **2022**(1), 32 (2022)
2. K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using fourier parameters. IEEE Trans. Affect. Comput. **6**(1), 69–75 (2015)
3. D. Tang, P. Kuppens, L. Geurts, T. van Waterschoot, End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network. EURASIP J. Audio Speech Music Process. **2021**(1), 18 (2021)
4. L. Sun, S. Fu, F. Wang, Decision tree svm model with fisher feature selection for speech emotion recognition. EURASIP J. Audio Speech Music Process. **2019**(1), 1–14 (2019)
5. P. Ekman, An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
6. J.A. Russell, A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
7. A. Cabri, F. Masulli, Z. Mnasri, S. Rovetta et al., Emotion recognition from speech: an unsupervised learning approach. Int. J. Comput. Intell. Syst. **14**(1), 23 (2020)
8. K.R. Scherer et al., On the nature and function of emotion: a component process approach. Approaches Emot. **2293**(317), 31 (1984)
9. Rao, K.S., Koolagudi, S.G. Robust emotion recognition using spectral and prosodic features. In: Springer Science & Business Media, Springer, NewYork (2013)

10. Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech emotion classification using attention-based lstm. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(11), 1675–1685 (2019)
11. Y. Xu, W. Wang, H. Cui, M. Xu, M. Li, Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. EURASIP J. Audio Speech Music Process. **2022**(1), 1–16 (2022)
12. E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, A. Stolcke, Modeling prosodic feature sequences for speaker recognition. Speech Commun. **46**(3–4), 455–472 (2005)
13. S.R. Kshirsagar, T.H. Falk, Quality-aware bag of modulation spectrum features for robust speech emotion recognition. IEEE Trans. Affect. Comput. **13**(4), 1892–1905 (2022)
14. M. Geravanchizadeh, E. Forouhandeh, M. Bashirpour, Feature compensation based on the normalization of vocal tract length for the improvement of emotion-affected speech recognition. EURASIP J. Audio Speech Music Process. **2021**, 1–19 (2021)
15. J.L. Jacobson, D.C. Boersma, R.B. Fields, K.L. Olson, Paralinguistic features of adult speech to infants and small children. Child Dev. **54**(2), 436–442 (1983)
16. S.M. Tsai, in *2013 1st International Conference on Orange Technologies (ICOT)*, A robust zero-watermarking algorithm for audio based on LPCC (IEEE, 2013), pp. 63–66
17. C. Ittichaichareon, S. Suksri, T. Yingthawornsuk, in *International conference on computer graphics, simulation and modeling (ICGSM'2012)*, vol. 9, Speech recognition using mfcc, Pattaya, Thailand (2012)
18. T.L. Nwe, S.W. Foo, L.C. De Silva, Speech emotion recognition using hidden markov models. Speech Commun. **41**(4), 603–623 (2003)
19. F. Albu, D. Hagiescu, L. Vladutu, M.A. Puica, in *EDULEARN15 Proceedings*, Neural network approaches for children's emotion recognition in intelligent learning applications (IATED, 2015), pp. 3229–3239
20. G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
21. L.E. Peterson, K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
22. B. Schuller, G. Rigoll, M. Lang, in *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 1, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture (IEEE, 2004), pp. I–577
23. K. Han, D. Yu, I. Tashev, in *INTERSPEECH 2014*, Speech emotion recognition using deep neural network and extreme learning machine, ISCA, Singapore (2014)
24. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. Narayanan, in *Proc. INTERSPEECH 2010,* The INTERSPEECH 2010 paralinguistic challenge, ISCA, Makuhari, Japan, (2010), pp. 2794–2797
25. F. Eyben, M. Wöllmer, B. Schuller, in *Proceedings of the 18th ACM international conference on Multimedia*, Opensmile: the munich versatile and fast open-source audio feature extractor, ACM, New York, United States (2010), pp. 1459–1462
26. J. Joy, A. Kannan, S. Ram, S. Rama, Speech emotion recognition using neural network and MLP classifier, International Journal of Engineering Science and Computing, Pearl Media Publications PVT LTD, **10**(4), pp. 25170–25172 (2020)
27. B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, in *Proceedings of the 14th python in science conference*, vol. 8, librosa: audio and music signal analysis in python (Citeseer, 2015), pp. 18–25
28. F. Albu, A. Mateescu, N. Dumitriu, in *International Conference on Microelectronics and Computer Science*, Architecture selection for a multilayer feedforward network (Citeseer, 1997), pp. 131–134
29. C. Xiang, S.Q. Ding, T.H. Lee, Geometrical interpretation and architecture selection of MLP. IEEE Trans. Neural Netw. **16**(1), 84–96 (2005)
30. T. Andersen, T. Martinez, in *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, vol. 3, Cross validation and MLP architecture selection (IEEE, 1999), pp. 1614–1619
31. G. Roffo, Feature selection library (matlab toolbox). arXiv preprint arXiv:1607.01327 (2016)
32. S. Russell, P. Norvig, Artificial intelligence: a modern approach, Prentice Hall, London, United Kingdom (2003)
33. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
34. H. Liu, H. Motoda, in *Chapman & Hall/CRC*, Computational methods of feature selection (Chapman & Hall/CRC data mining and knowledge discovery series), Chapman and Hall/CRC, Florida, United States (2007)
35. H. Zeng, Y.m. Cheung, Feature selection and kernel learning for local learning-based clustering. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1532–1547 (2010)
36. Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, in *Twenty-second international joint conference on artificial intelligence*, L2, 1-norm regularized discriminative feature selection for unsupervised, AAAI Press, Washington, United States (2011)
37. F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, et al., in *INTERSPEECH*, vol. 5, A database of German emotional speech. ISCA, Lisbon, Portugal (2005), pp. 1517–1520
38. Y. Fu, X. Yuan, in *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*, Composite feature extraction for speech emotion recognition (IEEE, 2020), pp. 72–77
39. W.A. Jassim, R. Paramesran, N. Harte, Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features. IET Signal Proc. **11**(5), 587–595 (2017)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.