BOHR International Journal of Internet of things,
Artificial Intelligence and Machine Learning
2023, Vol. 2, No. 1, pp. 1–10
DOI: 10.54646/bijiam.2023.11
www.bohrpub.com

METHODS

Emotion recognition based on speech signals by combining empirical mode decomposition and deep neural network

Shing-Tai Pan[1]*, Ching-Fa Chen[2] and Chuan-Cheng Hong[3]

[1]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, R.O.C
[2]Department of Electronic Engineering, Kao Yuan University, Kaohsiung, Taiwan, R.O.C
[3]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, R.O.C

***Correspondence:**
Shing-Tai Pan,
stpan@nu.edu.tw

Received: 05 January 2023; **Accepted:** 17 January 2023; **Published:** 30 January 2023

This paper proposes a novel method for speech emotion recognition. Empirical mode decomposition (EMD) is applied in this paper for the extraction of emotional features from speeches, and a deep neural network (DNN) is used to classify speech emotions. This paper enhances the emotional components in speech signals by using EMD with acoustic feature Mel-Scale Frequency Cepstral Coefficients (MFCCs) to improve the recognition rates of emotions from speeches using the classifier DNN. In this paper, EMD is first used to decompose the speech signals, which contain emotional components into multiple intrinsic mode functions (IMFs), and then emotional features are derived from the IMFs and are calculated using MFCC. Then, the emotional features are used to train the DNN model. Finally, a trained model that could recognize the emotional signals is then used to identify emotions in speeches. Experimental results reveal that the proposed method is effective.

Keywords: speech emotion recognition, empirical mode decomposition, deep neural network, mel-scale frequency cepstral coefficients, hidden markov model

1. Introduction

People can, most of the time, sense precisely the emotion of a speaker during communication. For example, people can detect an angry emotion from a loud harsh voice and a happy emotion from a voice full of laughter. This means that people can easily get information about the mood of a person simply by listening to them. In fact, emotion is a piece of vital information that speech signals carry apart from the verbal corpus (1). Human-computer interface (HCI) that can automatically detect emotions from the speech is then reasonable and promising. Recently, studies concerning automatic emotion recognition from speech attract a lot of attention. These studies include topics across many fields, for example, psychology, sociology, biomedical science, and education. All these foci on the impact of emotion on their health, and how to recognize the status of spirit one is from his speech. Speech is the most important media of these studies due to the following reasons: (a) the availability of fast computing systems, (b) the effectiveness of various signal processing algorithms, and (c) the acoustic differences in speech signals that are naturally embedded in various emotional situations (2).

Computing ability has enormously improved in this decade. Hence, it becomes possible and innovative to develop a system with machine learning methods or deep learning methods, which can recognize automatically people's emotions from their speeches. There are some literatures on this topic. For example, refer to the papers (1–6). In the paper (1), the Fuzzy Rank-Based Ensemble of Transfer Learning Model is used for speech emotion recognition. In the paper (2), the empirical mode decomposition (EMD) is applied to decompose speech signals and obtain non-linear features for emotion recognition. This paper (3) uses Hidden Markov Model (HMM) for speech emotion recognition. A hybrid system

of using signals from faces and voices to recognize people's emotions is proposed in the paper (4). An exploration of various models and speech features for speech emotion recognition is introduced in papers (5, 6). However, according to the studies (7, 8), the main topics of automatic emotion recognition from speeches include the selection of a database, feature extraction problems, and development of recognition algorithms (8). The exploration of the recognition algorithm is an important issue in the emotion recognition problem. There are some algorithms used for this application, for example, HMM (9), Support Vector Machine (SVM) (10, 11), Gaussian Mixture Model (GMM) (12), K-Nearest Neighbors (KNN) (13), and Artificial Neural Network (ANN) (14). A method of combining speech and image for emotion recognition is explored in (15). However, this method will take more computation time, and more hardware resources are required. Hence, this paper focuses on emotion recognition based on speech signals. Since ANN mimics the architecture of neurons in the organism to process signals, it has some advantages over the other methods: excellent fault tolerance capacity, good learning ability, and suitable for nonlinear regression problems. Hence, this paper adopts ANN as an emotion recognition algorithm. A supervised ANN, Deep Neural Network (DNN), is used in this paper for training the emotion model of speeches and then recognizing emotions from speeches.

The objective of this paper was to improve the emotion recognition rates based on speech signals by applying deep learning methods due to the massive progress in the capability of deep learning methods in recent years. A DNN will be adopted in this paper for this purpose. First, this paper applies the EMD method to improve emotional feature extraction. The weighted IMFs decomposed by using EMD will be summed to obtain the emotional features from speeches. The weights for the IMFs are designed by using genetic algorithms. The weighted sum of IMFs is then calculated to extract MFCC features. The MFCCs are then used to train classifiers for emotion recognition. As to the classifier, since HMM has been used for decades for speech recognition and has been successfully applied in many applications of speech recognition, this paper will use the emotion recognition results from HMM for comparison. Besides, for the purpose of saving computation time and using fewer hardware resources, the DNN architecture will be designed to be as simple as possible to achieve better emotion recognition rates compared with those obtained by using HMM.

The organization of this paper is as follows. For readability, Section 2 will briefly introduce the preprocessing and feature extraction methods for speeches in this paper. The EMD method used for extracting emotional features is introduced in Section 3. The classifiers HMM and DNN are then introduced in Section 4 and Section 5, respectively. The experimental results of emotion recognition using the proposed methods are revealed in Section 6. Finally, Section 7 makes some conclusions for this paper.

2. Preprocessing and feature extraction for speech

2.1. Framing speech

Speech signals are non-stationary signals and vary with time. It is necessary for speech signal processing to divide a speech into several short blocks to get more stationary signals. Hence, frames are taken from speech signals at the first step. The extracted frames are always overlapped to make the frame contain some previous information. Different rates of overlap will make difference to the features of speech signals. However, some experiments are required to choose a suitable over-lapping rate. A frame with 256 points is used in this paper. That is, for a speech with 8 kHz sampling rates and 1 second length of time, there will be about 32 frames obtained from framing the speech. In this paper, uniform sampling of speech signals is used since it is more robust and less biased than non-uniform sampling (16). However, since emotional speeches are always different in length of time, there are different numbers of frames after framing different speeches.

2.2. Speech preemphasis

While speeches transmit in the air, high-frequency signals in the speeches are attenuated more than low-frequency signals in speeches. Hence, a high-pass finite-impulse-response (FIR) filter is applied to speech signals to enhance the high-frequency components. A high-pass filter can be described as follows (8):

$$S_{pe}(n) = S_{of}(n) - 0.97 \times S_{of}(n-1),\ 1 \le n \le N \quad (1)$$

in which $S_{pe}(n)$ is the output of the FIR filter; $S_{of}(n)$ is the original speech signal; and N is the number of points in a frame.

2.3. Applying hamming window

Fourier transform is used commonly to calculate features of the speeches. However, due to the discontinuity at the start and at the end of a frame, high-frequency noisy signals may occur when the Fourier transform is taken on the frame. To solve this problem, a hamming window will be applied to the frames to reduce the effects caused by the noises. The hamming window is described by the following equation (9):

$$W(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2n\pi}{N-1}),\ 0 \le n \le N-1; \\ 0,\ otherwise \end{cases} \quad (2)$$

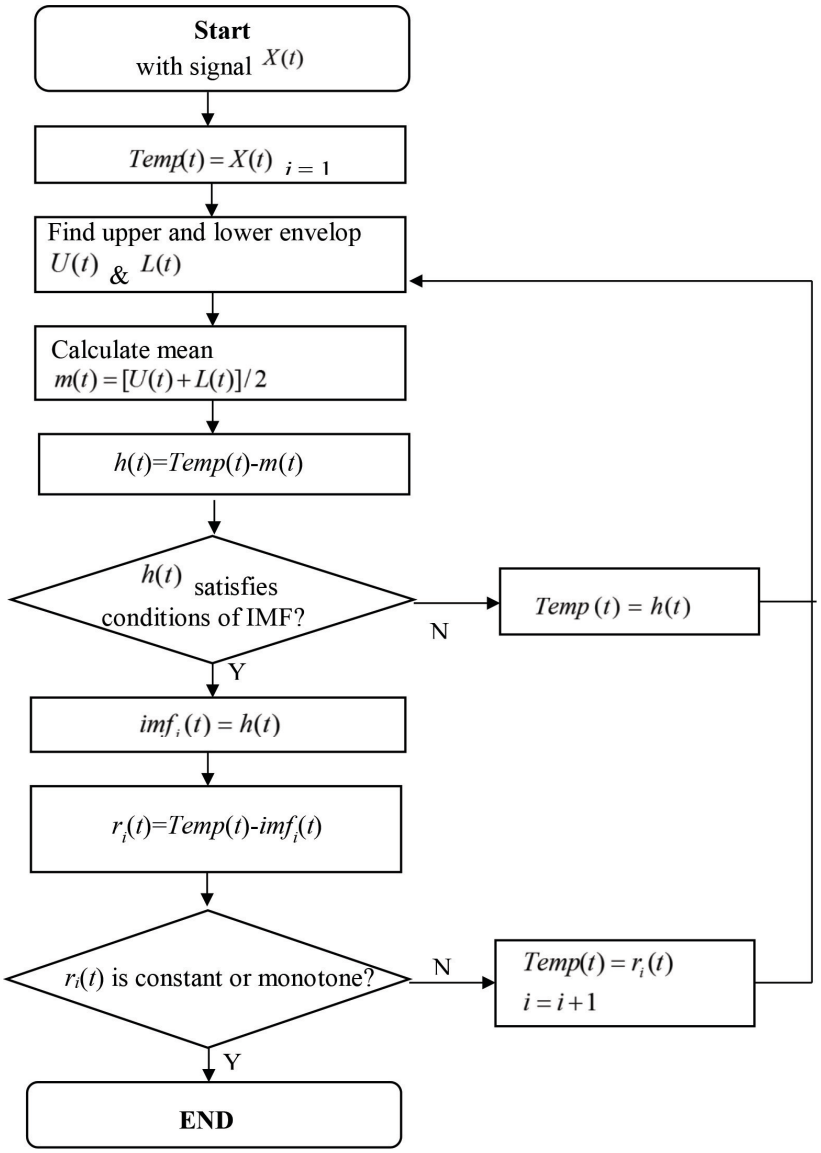

FIGURE 1 | Flowchart of empirical mode decomposition (EMD).

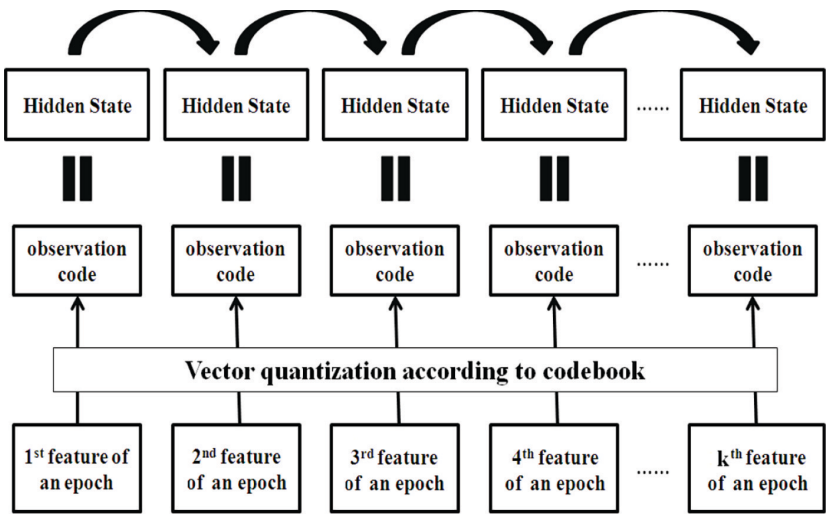

FIGURE 2 | Mechanism of Hidden Markov Model (HMM).

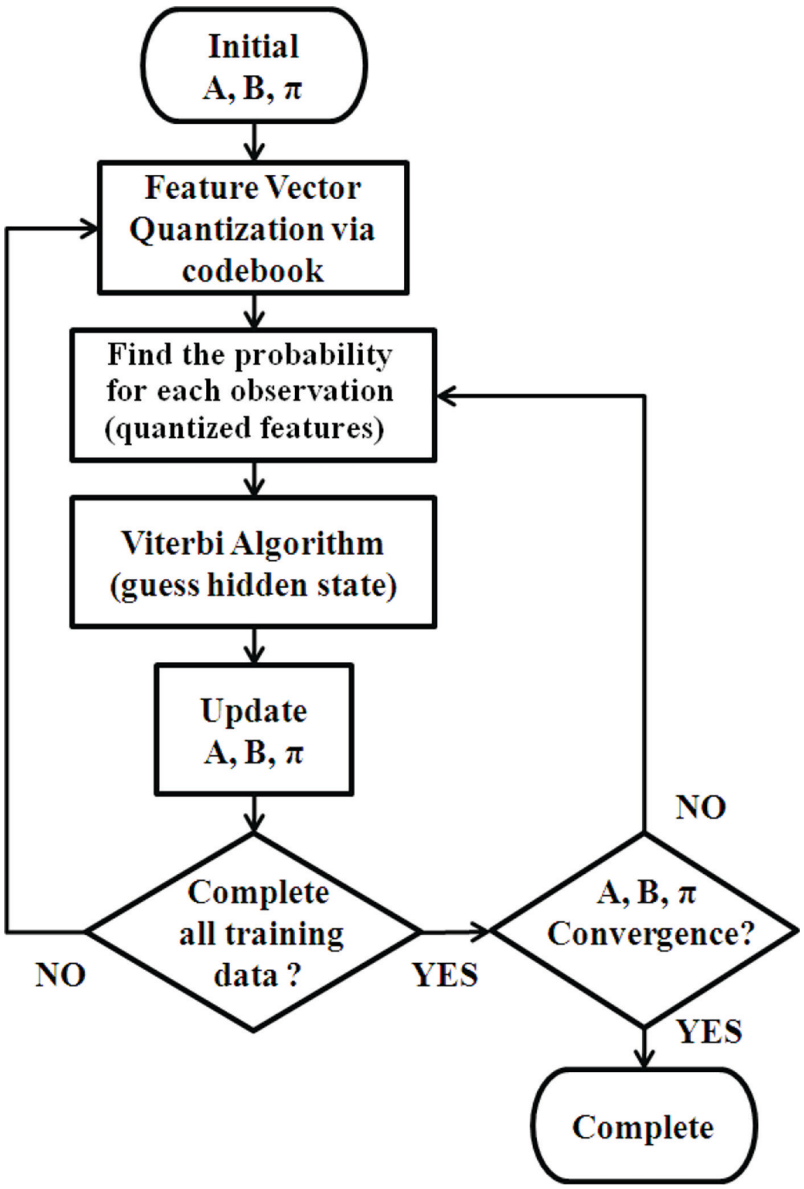

FIGURE 3 | Flowchart for the training of HMM.

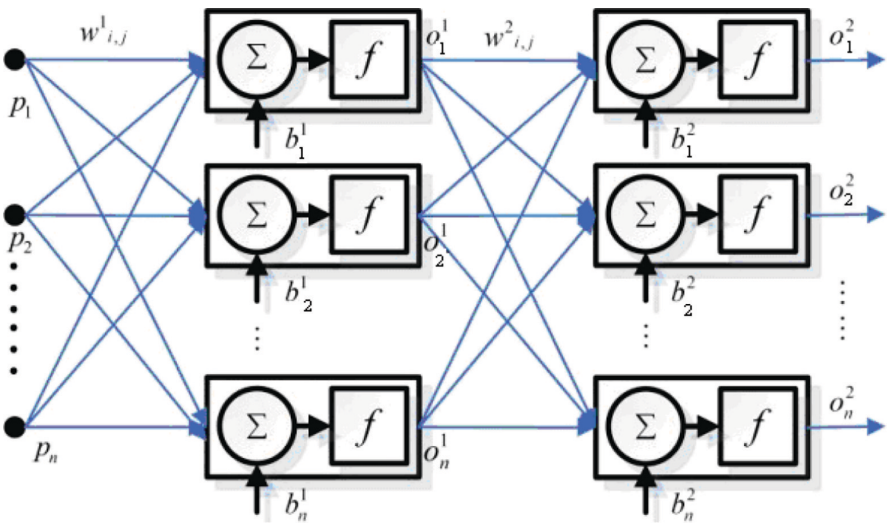

FIGURE 4 | The structure of the artificial neural network (ANN).

TABLE 1 | Description of Berlin emotional database.

Name of database		Berlin emotional database		
Language		German		
Speaker		Gender	Number	Total
		Male	5	10
		Female	5	
Sentence		Type	Number	Total
		Long sentence	5	10
		Short sentence	5	
Emotion		Type		Number
		anger		7
		joy		
		sadness		
		fear		
		disgust		
		boredom		
		neutral		
Facility for recording		Sennheiser MKH 40-P48 microphone		
		Tascam DA-P1 portable DAT recorder		
File	Data acquisition	Sampling rate: 16 kHz		
		Resolution: 16bits		
		Channel: Mono		
		Format: wav		
	Number	535		

where N is the number of points in a frame. And,

$$F(n) = W(n) \times S(n) \tag{3}$$

in which $S(n)$ is the nth point in a frame, and $F(n)$ is the result signal after applying a hamming window to the frame.

2.4. Fast fourier transform

To calculate Mel-Frequency Cepstral Coefficients (MFCC) for a frame, the speech signals will be presented in the frequency domain. Since the speech signals are presented initially in the time domain, fast Fourier transform (FFT) will be applied to the frames to transform them into frequency-domain representation. FFT can be described as follows (10):

$$X_k = \sum_{n=0}^{N-1} S_w(n) \times W_N^{kn},\ 0 \le k \le N-1 \tag{4}$$

$$W_N = e^{\frac{-j2\pi}{N}} \tag{5}$$

2.5. Mel-frequency cepstral coefficients

The feature Mel-frequency cepstrum simulates the reception properties of human ears. The MFCC will be calculated for each frame. To calculate MFCC, FFT is applied to speech frames first. Then, Mel triangular band-pass filter is applied to the results of the FFT, $X(k)$. The Mel triangular band-pass filter is described by the following equation:

$$B_m(k) = \begin{cases} 0, k < f_{m-1} \\ \frac{k-f_{m-1}}{f_m - f_{m-1}}, f_{m-1} \le k \le f_m \\ \frac{f_{m+1}-k}{f_{m+1}-f_m}, f_m \le k \le f_{m+1} \\ 0, f_{m+1} < k \end{cases} \tag{6}$$

where M denotes the number of filters and $1 \le m \le M$. The logarithm is then taken on the summation of the product of the frequency representation $X(k)$ and Mel triangular band-pass filter $B_m(k)$ as follows:

$$Y(m) = \log\left\{\sum_{k=f_{m-1}}^{f_{m+1}} |X_k|\, B_m(k)\right\}. \tag{7}$$

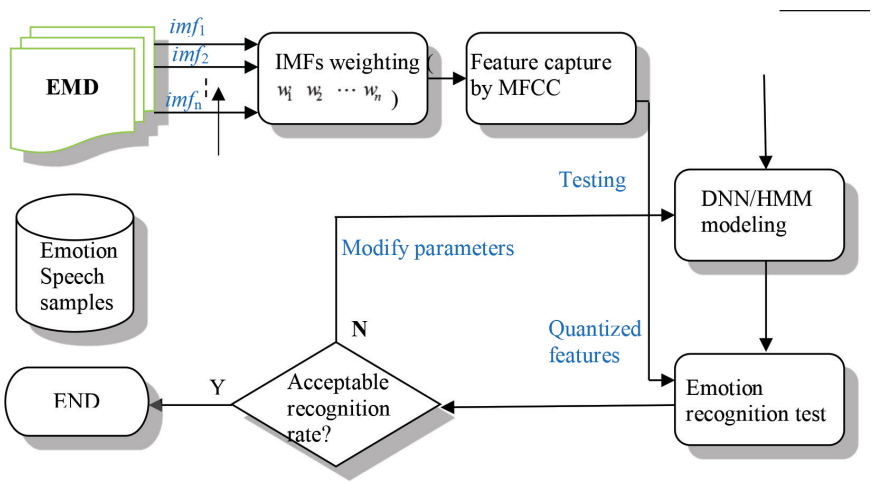

FIGURE 5 | The steps of the experiments.

TABLE 2 | Parameters and activation function for deep neural network (DNN).

Model	DNN
No. of hidden layers	5
No. of neuron in hidden layers	13
Activation function	Hyperbolic tangent function
Iterations	100000
Goal of error	10^(-15)
Learning rate	0.1

Then, the discrete cosine transform is applied to $Y(m)$ as follows:

$$c_x(n) = \frac{1}{M}\sum_{m=1}^{M} Y(m) \times \cos\left(\frac{\pi n(m-\frac{1}{2})}{M}\right) \quad (8)$$

in which cx (n) is MFCC. In this paper, the first 13 coefficients of cx (n) are calculated and then formed as a feature vector. These MFCCs for the frames of training speeches are used to train DNN, and those of testing speeches are used to be tested with the trained DNN.

3. Empirical mode decomposition

In this paper, EMD is used to decompose emotional speech signals into various emotional components, which are defined as intrinsic mode functions (IMFs). An IMF must satisfy the following two conditions (17):

(1) The number of local extremes and the number of zero-crossings differ at most by one.

(2) Upper and lower envelopes of the function are symmetric.

The steps of EMD are described as follows. It is noticed that, in this paper, the Cubic Spline (17) is adopted to construct the upper envelop and lower envelop of the signals in the process of deriving IMFs. Let the original signal be $_{X(t)}$ and $Temp(t) = X(t)$.

Step 1:

Find the upper envelope $U(t)$ and lower envelope $L(t)$ of the signal $Temp(t)$. Calculate the mean of the two envelops $m(t) = [U(t) + L(t)]/2$. The intermediate signal

TABLE 3 | Recognition rates using DNN without empirical mode decomposition (EMD) for 7 emotions based on 10-fold validation.

Emotions Experiments	Anger (%)	Joy (%)	Sadness (%)	Fear (%)	Disgust (%)	Boredom (%)	Neutral (%)
1	65.00	70.00	85.00	80.00	95.00	40.00	45.00
2	90.00	55.00	80.00	65.00	90.00	75.00	65.00
3	75.00	75.00	90.00	60.00	95.00	45.00	40.00
4	60.00	50.00	80.00	10.00	100.00	15.00	60.00
5	0	90.00	0	85.00	100.00	100.00	55.00
6	75.00	70.00	90.00	35.00	95.00	20.00	28.57
7	65.00	35.00	70.00	10.00	100.00	25.00	25.00
8	70.00	70.00	100.00	75.00	100.00	60.00	65.00
9	75.00	70.00	90.00	70.00	95.00	75.00	60.00
10	80.00	75.00	85.00	75.00	100.00	55.00	60.00
Avg.	65.50	66.00	77.00	56.50	97.00	51.00	50.36

TABLE 4 | Recognition rates using DNN *with* EMD for 7 emotions based on 10-fold validation.

Emotions Experiments	Anger (%)	Joy (%)	Sadness (%)	Fear (%)	Disgust (%)	Boredom (%)	Neutral (%)
1	55.00	80.00	80.00	80.00	95.00	60.00	40.00
2	85.00	80.00	80.00	60.00	90.00	75.00	70.00
3	90.00	70.00	80.00	60.00	95.00	50.00	45.00
4	80.00	70.00	90.00	20.00	100.00	20.00	60.00
5	0	75.00	0	85.00	100.00	100.00	60.00
6	65.00	70.00	85.00	25.00	95.00	25.00	28.57
7	65.00	50.00	65.00	10.00	100.00	15.00	45.00
8	75.00	75.00	100.00	70.00	95.00	70.00	70.00
9	85.00	75.00	90.00	70.00	95.00	75.00	55.00
10	95.00	85.00	95.00	85.00	100.00	60.00	65.00
Avg.	69.50	73.00	76.50	56.50	96.50	55.00	53.86

$h(t)$ is calculated as follows:

$$h(t) = Temp(t) - m(t) \quad (9)$$

Step 2:

Check whether the intermediate signal $h(t)$ satisfies the conditions of IMF or not. If it does, then the first IMF is obtained as follows: $imf_1(t) = h(t)$, and we moved to the next step or assigned the intermediate signal $h(t)$ as $Temp(t)$ and moved back to Step 1.

Step 3:

Calculate the residue $r_1(t)$ as follows:

$$r_1(t) = Temp(t) - imf_1(t). \quad (10)$$

Assign the signal $r_1(t)$ as $X(t)$ and repeat Step 1 and Step 2 to find imf_2 (t).

Step 4:

Repeat Step 1 to Step 3 to find the subsequent IMFs as follows:

$$r_n(t) = r_{n-1}(t) - imf_n(t), n = 2, 3, 4, \ldots. \quad (11)$$

If the signal $_{r_n(t)}$ is constant or a monotone function, then the EMD procedure is completed. Then, the following decomposition of $_{X(t)}$ is obtained as follows:

$$X(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t) \quad (12)$$

The flowchart of EMD is depicted in **Figure 1**. In this paper, a weighted sum of *imf*s is proposed to recover the emotional components and is written as the following equation (13). The value of weights w_i will be set according to the results in (7).

$$X(t) \sum_{i\ 1}^{n} w_i \cdot imf_i(t) \quad (13)$$

4. Hidden markov model

In this paper, a discrete HMM is used as a comparison with the proposed DNN method. The feature MFCCs that are extracted from the speech signals after EMD processing are used to train HMM and then for testing. The MFCC features of speech signals are arranged as a time series according to the order of frames obtained from framing each speech signal. The time series of MFCCs is treated as the observation of the HMM model, and the hidden states of the model will be estimated using the Viterbi Algorithm (18). **Figure 2** shows the mechanism of the HMM model with the features, observations, and hidden states. The parameters in HMM $\lambda = (A, B, \pi)$ are explained as follows (18–20):

$A = [a_{ij}]$, $a_{ij} = P(q_t = x_j | q_{t-1} = x_i)$, the probability of hidden state x_i transfers to hidden state x_j,

$B = [b_j(k)]$, $b_j(k) = P(o_t = v_k | q_t = x_j)$, the probability of *kth* observation v_k happens at the *j*th hidden state x_j,

$\pi = [\pi_i]$, $\pi_i = P(q_1 = x_i)$, the probability of hidden state $_{x_i}$ happens at the initial of the time series,

$X = (x_1, x_2, \cdots, x_N)$, is the hidden state of HMM.

The training process for modeling HMM is depicted in **Figure 3**. The initial values of matrices A, B, and π are given randomly. A trained codebook is then used to quantize MFCC features. The matrices A, B, and π are then updated using the Viterbi Algorithm (18). This process will repeat until the parameters in A, B, and π converge. Then, the training process for HMM is completed.

5. Deep neural network

The architecture of ANN is constructed based on connections of the multiple-layer perceptron. Each layer comprises several neurons. The transition of signals between neurons is similar to that between neurons in an organism. In this paper, a deep neural network structure, i.e., the network with several hidden layers, is proposed and will be

TABLE 5 | Comparison of recognition rates for 10-fold experiments using DNN with and without EMD.

	Set 1 (%)	Set 2 (%)	Set 3 (%)	Set 4 (%)	Set 5 (%)	Set 6 (%)	Set 7 (%)	Set 8 (%)	Set 9 (%)	Set 10 (%)	Avg. (%)
Without EMD	68.571	74.286	68.571	53.571	61.429	58.865	47.143	77.143	76.429	75.714	66.172
With EMD	70.0	77.143	70.0	62.857	60.0	56.028	50.0	79.286	77.857	83.571	68.674

TABLE 6 | Comparison of recognition rates for 7 emotions using DNN with and without EMD.

	Anger (%)	Joy (%)	Sadness (%)	Fear (%)	Disgust (%)	Boredom (%)	Neutral (%)	Avg. (%)
without EMD	65.50	66.00	77.00	56.50	97.00	51.00	50.36	66.19
With EMD	69.50	73.00	76.50	56.50	96.50	55.00	53.86	68.69

Moreover, the results gained by the proposed method are compared with those in ref. (9), in which HMM was used. **Table 7** shows the results of the comparison. It is obvious that the proposed method has higher recognition rates at each fold of experiments as well as at the average recognition rates of the 10-fold experiments in both cases whether EMD is used or not. This verifies the performance of the proposed method.

compared with HMM. This structure allows us to train the neural network deeply. The structure of a three-layer ANN is shown in **Figure 4** (21).

In **Figure 4**, P_n is nth input, $w_{i,j}^k$ is the weight between the ith and jth neuron in $(k\text{-}1)^{th}$ layer and k^{th} layer, respectively; o_n^k is nth output in the kth layer, and b_n^k is the bias of nth neuron in the k^{th} layer.

After completing the calculation of MFCC for all emotional speech signals, we started to train ANN using the MFCCs obtained from training speech signals. The MFCC with a label of emotion will be fed to the input of ANN, and then the output of the ANN is used to compute the errors with respect to the label of MFCC in the input. The output function of ANN is described in equations (14) and (15).

$$o_n^k = f\left(\sum_{\forall i} w_{i,n}^k o_i^{k-1} - b_n^k\right) \tag{14}$$

$$f(x) = \tanh(x) \tag{15}$$

The back-propagation algorithm with the steepest descent method (SDM) is used to train ANN by updating the weights based on the error function between the output and the goal to find the optimal parameters of ANN. The back-propagation algorithm is described in equations (16)-(21).

$$\delta_n = (T_n - o_n^2) \times f'\left(\sum_i w_{i,n}^2 o_i^1 - b_n^2\right) \tag{16}$$

$$\Delta W_{i,n}^2 = \eta \times \delta_n \times o_i^1 \tag{17}$$

$$\Delta b_n^2 = -\eta \times \delta_n \tag{18}$$

$$\delta_i = \left(\sum_n \delta_n w_{i,n}^2\right) \times f'\left(\sum w_{r,i}^1 P_r - b_i^1\right) \tag{19}$$

$$\Delta W_{r,i}^1 = \eta \times \delta_i \times P_r \tag{20}$$

$$\Delta b_i^1 = -\eta \times \delta_i. \tag{21}$$

where T_n is the goal of the output of ANN; η is the learning step.

TABLE 7 | Comparison of the results by the proposed method and those by the Hidden Markov Model (HMM) (9)

	HMM (%)	HMM+EMD (%)	DNN (%)	DNN +EMD (%)
1	57.14	62.86	68.571	70.0
2	56.43	67.14	74.286	77.143
3	55.71	67.86	68.571	70.0
4	47.86	54.29	53.571	62.857
5	66.43	76.43	61.429	60.0
6	47.14	58.57	58.865	56.028
7	39.29	52.86	47.143	50.0
8	64.29	74.29	77.143	79.286
9	69.29	73.57	76.429	77.857
10	65.71	77.14	75.714	83.571
Avg.	56.93	66.50	66.172	68.674

6. Experimental results

The experiments conducted in this paper are performed on a personal computer (PC), and the algorithms for the experiments are implemented using MATLAB. The emotional speech database used for the experiments in this paper is the Berlin emotional database (22). The database is recorded in the German language by 10 professional actors. The professional actors include 5 males and 5 females. All the speeches in this database are sampled with 8 kHz of 16bits length in.wav format. The details of the Berlin emotional database are described in **Table 1**. A 10-fold cross-validation method is adopted for this experiment.

The experiments conducted in this paper are performed by the following steps:

1. Classify the dataset from Berlin Emotional Database into the training dataset and testing dataset.
2. Separate all the speeches in the training dataset and testing dataset into various IMFs and recombine the IMFs with the weights in ref. (9).
3. Calculate MFCC for the results in Step 2.
4. Train the DNN model and the HMM model using the feature MFCC obtained in Step 3.
5. Repeat Step 1 to Step 4 until the recognition rate meets the goal set in this experiment.
6. Use the model trained in Step 5 for testing.

The steps of performing the experiments in this paper are described in **Figure 5**.

In this experiment, the structure of DNN, activation functions, and settings for the experiment are described in **Table 2**. The number of hidden layers is set to 5 to fulfill a deep learning architecture. A commonly used activation, hyperbolic tangent function, is adopted here.

The results of the experiments with and without EMD will be compared to verify the advantage of applying EMD in the experiments. First, the experimental numerical results without EMD for the 7 emotions in the Berlin database with 10-fold validation are shown in **Table 3**. The average recognition rates of the 7 emotions are from 50.36% to 97%. The recognition rates for some emotions are high, especially the emotions "disgust" and "sadness". This is because the features of these two emotional speeches are distinct while the emotions "fear", "boredom," and "neutral" have similar

features to each other. Hence, the recognition rates are relatively low. Then, EMD is applied for emotion extractions of speeches. The experimental results for the 7 emotions in the Berlin database with 10-fold validations are shown in **Table 4**.

The comparisons of the recognition rates between the DNN with and without EMD for 10-fold experiments are shown in **Table 5**. It can be seen, from the red context in the table, that the recognition rates of most runs and the average recognition rate in the experiment are better when EMD is applied. Moreover, according to **Table 6**, when EMD is applied for extractions of emotion components, better recognition rates are gained for emotions "anger", "joy", "boredom," and "neutral". The recognition rate for the emotion "fear" remains the same. The emotions "sadness" and "disgust" have slightly lower recognition rates. The average recognition rate is better than that without using EMD. Please refer to the red context in **Table 6** for more details. These experimental results verify the advantage of using EMD to extract emotional components.

7. Conclusion

In this paper, EMD is applied to extract the emotional features from speeches. The experimental results of this paper reveal that the emotion recognition rates are better for both classifiers, i.e., HMM and DNN, after applying EMD for emotional feature extractions. However, according to **Table 6**, EMD does not work well for two emotions, i.e., "sadness" and "disgust". It is likely that the features of the two emotions are similar, and EMD cannot effectively distinguish them. In the future work, some advanced EMD, such as Ensemble EMD, may be used to get a better extraction of emotional features from emotional speeches and hence get better emotion recognition rates. Besides, in this paper, a simple DNN is designed to get better recognition rates than those gained by using HMM in both cases whether EMD is applied or not. According to **Table 7**, the improved recognition rates are about 10% and 2% respective to that the EMD is not and is applied to speech signals. However, in our experiments, only a few minutes are needed to train HMM while DNN used in this paper takes more than 40 minutes. Consequently, the improvement of time consumption of DNN is still an open problem.

Author contributions

S-TP: Conceptualization, methodology, investigation, and writing—original draft preparation. C-FC: validation, formal analysis, and writing—review and editing. C-CH: software and resources.

Funding

This research was funded by the Ministry of Science and Technology of the Republic of China, grant number MOST 109-2221-E-390-014-MY2. This research work was supported by the Ministry of Science and Technology of the Republic of China under contract MOST 108-2221-E-390-018.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Sahoo KK, Dutta I, Ijaz MF, Wozniak M, Singh PK. TLEFuzzyNet: fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches. *IEEE Access*. (2021) 9:166518–30.
2. Krishnan PT, Joseph Raj AN, Rajangam V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intelligent Syst*. (2021) 7:1919–34.
3. Schuller B, Rigoll G, Lang M. Hidden markov model-based speech emotion recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*. Baltimore, MD: IEEE (2003). p. 1-4.
4. Fragopanagos N, Taylor JG. Emotion recognition in human-computer interaction. *Neural Networks*. (2005). 18:389–405.
5. Cen L, Ser,W, Yu ZL, Cen W. Automatic recognition of emotional states from human speeches. In: A. Herout, editor. *Pattern recognition Recent Advances*. London: intechopen. (2011).
6. Wu S, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* (2011) 53:768–85.
7. Picard RW. *Affective Computing*. Cambridge, MA: MIT press (1997).
8. Basu S., Chakraborty, J., Bag A., Aftabuddin, M. A review on emotion recognition using speech. In: *Proceedings of the International Conference on Inventive Communication and Computational Technologies (ICICCT)*. Coimbatore (2017). p. 109–14.
9. Lee YW, Pan ST. *Applications of Empirical Mode Decomposition on the Computation of Emotional Speech Features*. Taiwan: National University of Kaohsiung (2012).
10. Zhu L, Chen L, Zhao D, Zhou J, Zhang W. Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN. *Sensors*. (2017) 17:1694.
11. Trabelsi I, Bouhlel MS. Feature selection for GUMI kernel-based SVM in speech emotion recognition. In: Information Reso Management Association, editor. *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications*. Pennsylvania: IGI Global (2017). p. 941–53.
12. Patel P, Chaudhari A, Kale R, Pund M. Emotion recognition from speech with gaussian mixture models & via boosted GMM. *Int J Res Sci Eng*. (2017) 3: 47–53.
13. Jo Y, Lee H, Cho A, Whang M. Emotion recognition through cardiovascular response in daily life using KNN classifier. In: J. J. Park, editor. *Advances in Computer Science and Ubiquitous Computing*. Singapore: Springer (2017). p. 1451–6.

14. Alhagry S, Fahmy AA, El-Khoribi RA. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *Emotion*. (2017) 8:355–8.
15. Ghaleb E, Popa M, Asteriadis S. Metric Learning-based multimodal audio-visual emotion recognition. *IEEE multimedia*. (2020) 27:1–8.
16. Shang Y. Subgraph robustness of complex networks under attacks. *IEEE Trans Syst Man Cybernet*. (2019) 49:821–32.
17. Huang NE. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*. (1996) 454:903–95.
18. Blunsom P. *Hidden Markov Model*. Parkville, VIC: The University of Melbourne (2004)
19. Pan ST, Hong TP. Robust speech recognition by DHMM with a codebook trained by genetic algorithm. *J Informat Hiding Multimedia Signal Processing*. (2012) 3:306–19.
20. Pan ST, Li WC. Fuzzy-HMM modeling for emotion detection using electrocardiogram signals. *Asian J Control*. (2020) 22:2206–16.
21. Pan ST, Lan ML. An efficient hybrid learning algorithm for neural network–based speech recognition systems on FPGA chip. *Neural Comput Appl*. (2014) 24:1879–85.
22. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. *Interspeech*. (2005) 5:1517–20.
23. Mathews JH, Fink KD. *Numerical Methods Using MATLAB*. 4th ed. Hoboken, NJ: Prentice-Hall (2004)