Frame Optimization in Speech Emotion Recognition Based on Improved EMD and SVM Algorithms

Chuan-Jie Guo[1], Shu-Ya Jin[1], Yu-Zhe Zhang[1], Chi-Yuan Ma[1], Muhammad Adeel[1], and Zhi-Yong Tao[1,2(✉)]

[1] Guangxi Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education, Guilin University of Electronic Technology, Guilin 541004, China
[2] Academy of Marine Information Technology, Guilin University of Electronic Technology, Beihai 536000, China
zytao@guet.edu.cn

Abstract. Emotional features of speech signals are one of the keys to human-computer interaction. However, there are still great difficulties and chances to extract emotional features. There is also great controversy regarding the part of signal preprocessing. This study divides the speech signal into small frames that overlap with a portion of the previous frame and adopts an improved empirical mode decomposition (EMD) based feature extraction method. The aim is to find the most suitable framing method. Each frame signal is processed by an improved EMD to generate a set of intrinsic mode functions (IMFs). Multidimensional features are extracted by calculating the central frequency and energy intensity of each IMF, and subsequently processing the center frequency of each IMF. Specifically, we focus on the top three IMFs in terms of energy intensity. Based on the improved algorithm, we investigate the effects of different frame lengths and frame shifts on the recognition rates of three emotion classifications: happy, angry, and sad. We find that the proposed method can reach the highest recognition rate when we use a 30 ms frame length with a 25% frame shift to separate the signals.

Keywords: Speech emotion recognition · Improved EMD · Signal framing · SVM

1 Introduction

Speech is one of the most common communication channels used to express human emotions. Emotions play an important role in human communication

Supported by Guangxi Natural Science Foundation (Grant Nos. 2021GXNSFDA075006, 2021GXNSFAA220086), National Natural Science Foundation of China (Grant No. 12064005), the Dean Project of Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology), and Innovation Project of GUET Graduate Education (2022YCXS045).

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 110–124, 2025.
https://doi.org/10.1007/978-3-031-60347-1_11

because of the rapid development of human-computer interaction and the wide usage of speech emotion recognition in various fields [1–5], such as automotive applications, aircraft piloting, medical services, and communication applications [6–8]. The successful detection of emotional states can help improve the efficiency of human-computer interaction [9–13]. In the domain of speech signal processing, improving the recognition performance of emotional speech signals is very important.

Over the past few decades, there has been rapid progress in the development of techniques used for extracting features from emotional speech signals. Commonly used methods are as follows. The main features extracted from speech emotion studies are rhythmic and phonological features. The features related to rhythm are the fundamental period, amplitude energy, speech rate, duration, etc. and their statistical values. Mel frequency cepstral coefficients (MFCCs) are used for recognition; if there is no interference, the recognition effect is very good, but in noise interference conditions, the recognition effect drops sharply. These methods require signal framing before use, but there is little literature discussing the impact of framing methods on feature extraction methods.

Due to the nonlinear and non-smooth nature of speech signals, some conventional methods have certain limitations and may only be suitable for specific speech samples. Huang et al. proposed the Hilbert-Huang transformation (HHT) [14,15], a time-frequency analysis technique for nonlinear and non-smooth signals. Various applications have demonstrated that it is better suited for analyzing nonlinear and non-smooth signals compared to conventional techniques.

HHT contains a method of adaptive time-frequency signal processing known as EMD [16,17]. It decomposes the signal from the timescale and generate a finite number of IMF components, which are local feature signals containing different timescales from the original signal. Because HHT has complete adaptivity, its EMD process does not need to set the basis functions in advance, which overcomes the shortcomings of wavelet analysis methods that rely on subjective experience and responds better to the physical properties of the signal. Therefore, we adopt an improved EMD feature extraction method to study frame optimization.

A major drawback of the EMD method is its tendency to exhibit mode mixing during the decomposition process, which can adversely affect the decomposition outcome. EMD has been continuously improved and has undergone several iterations, including EMD [16,17], ensemble empirical mode decomposition (EEMD) [18,19], complete ensemble empirical mode decomposition (CEEMD) [20,21], complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [22,23] and the improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) [24,25].

Compared with EMD, EEMD, CEEMD and CEEMDAN, the ICEEMDAN method is more advanced, and it is an excellent method for signal decomposition and noise reduction because it solves the problem of noise while avoiding mode mixing and also improves the decomposition efficiency and accuracy. In addition

to time-series signals, the ICEEMDAN method is suitable for analyzing and processing other types of signals, such as images, video, and audio.

The emotion database used in this research paper was developed by the PELL Lab for Chinese speech. It contains recordings of seven emotions expressed by two male and two female actors: angry, disgusted, fearful, happy, sad, surprised, and neutral. These pseudo-sentences have a sentence structure, but not the meaning of the speech itself, and the analysis of such speech data can effectively exclude the influence of sentence meaning on emotion feature extraction.

Based on the performance of the ICEEMDAN algorithm in removing mode mixing, we propose a central frequency projection (CFP)-based method for the combined classification of emotional speech signals [26]. The multidimensional features obtained using the proposed method can be used to distinguish emotional speech signals when a maximum IMF ranking process is introduced to further prevent pattern blending. In Sect. 2, we process the original signals according to different framing methods to construct a pool of data for classification, which is used to investigate the effect of framing on feature extraction.

In addition, in Sect. 3, we use the improved EMD method to decompose the signals within the data pool; features such as the energy intensity and center frequency of the decomposed IMFs are extracted. To extract multidimensional features, a method was proposed for rearranging the decomposed IMFs through IMF ordering, with the top three IMFs being selected. In Sect. 4, we develop a classification model and randomly build learning and testing datasets to conduct experiments to obtain the desired classification results, while we compare the sentiment recognition rates of the data pools built with different frame lengths and frame shifts in Sect. 2 to find the framing method with the highest recognition rate. Finally, in Sect. 5, we summarize our main findings regarding emotional speech signal recognition.

2 Data Pools of Emotional Speech Signals

The speech signals for the three different emotions used in this study were obtained from a database of Chinese speech emotional developed by the PELL lab [27]. This emotional speech database was designed to eliminate the influence of semantics, so that the impact of speech content on emotion could be excluded.

The database contains speech data of seven emotions-angry, disgusted, fearful, happy, sad, surprised and neutral-recorded by four actors in the absence of noise. We select three emotion signals between angry, happy and sad for one of the actors, and use our method for feature extraction and classification. Meanwhile, in order to ensure effectiveness, only one person's voice data is selected when classifying and recognizing emotions.

To study the effects of different frame lengths and frame shifts of signals on feature extraction, we split each type of emotional speech signal into five frames of lengths 15, 30, 45, 60, and 75 ms [28], with a frame shift equal to 25% of the frame length. We obtained five data pools of emotional speech signals, which are

used to study the effect of the frame length. After that, two more data pools with different frame shifts are obtained for the 30 ms frame length with 50% and 75% frame shifts, and these two data pools are compared with the 30 ms frame length with 25% frame shift to investigating the effect of frame shift.

3 ICEEMDAN Algorithm and Feature Extraction

To extract emotional features from speech signals, the first step is to process all signals in the data pool using ICEEMDAN. This method is used to minimize the mode mixing effect and extract all IMF components of each signal. Next, a maximum IMF ordering process is introduced. The center frequency of each IMF is obtained by averaging its associated instantaneous intensity and frequency. The center frequency was extracted by calculating the center frequency of each IMF and rearranging the IMF components based on energy. Using this method, the center frequencies of all frames in a sentence of speech can be obtained and processed to generate a set of multidimensional features.

3.1 ICEEMDAN Algorithm

The ICEEMDAN is an improved EMD algorithm. Let the original signal be denoted as $x(t)$, and the first IMF component obtained by decomposition of the EMD algorithm is expressed as

$$IMF_1(t) = x(t) - M(x(t)) \tag{1}$$

Therefore, the first IMF component obtained from the decomposition of the original $x(t)$ signal by the ICEEMDAN algorithm can be expressed as

$$IMF_1(t) = x(t) - \left\langle M\left(x^{(k)}(t)\right)\right\rangle = \left\langle E_1\left(x^{(k)}(t)\right)\right\rangle \tag{2}$$

In the above equation, $E(\cdot)$ is the IMF component, $M(\cdot)$ is the local mean, $\langle\cdot\rangle$ is the overall average, and $x^{(k)}(t)$ is a noise-added signal which can be expressed as

$$x^{(k)}(t) = x(t) + \alpha_0 E_1\left(w^{(k)}(t)\right) \tag{3}$$

In the above equation, $w^{(k)}(t)$ is the white noise, k=1,2,3,...,K denotes the number of times the Gaussian white noise is added, and α is the noise amplitude factor. Thus, the ICEEMDAN algorithm decomposition steps for the signal $x(t)$ are as follows.

(1) The first noise-added signal $x^{(k)}(t)$ is decomposed by the ICEEMDAN algorithm to produce the first residual component.

$$r_1(t) = \left\langle M\left(x^{(k)}(t)\right)\right\rangle \tag{4}$$

(2) Then the first IMF component is

$$IMF_1(t) = x(t) - r_1(t) \tag{5}$$

(3) In turn, the i-th residual component can be derived as

$$r_i(t) = \left\langle M\left(r_{i-1}(t) + \alpha_{i-1}E_i(w^{(k)}(t)\right)\right\rangle \tag{6}$$

(4) Then the i-th IMF component.

$$IMF_i(t) = r_{i-1}(t) - r_i(t) \tag{7}$$

(5) After obtaining all I IMF components, $x(t)$ can be expressed as

$$x(t) = \sum_{i=1}^{I} IMF_i(t) + r_I(t) \tag{8}$$

The above is the specific process of ICEEMDAN algorithm implementation, we process each frame signal in the data pool by ICEEMDAN algorithm to obtain their respective IMF components, which helps us in the subsequent feature extraction.

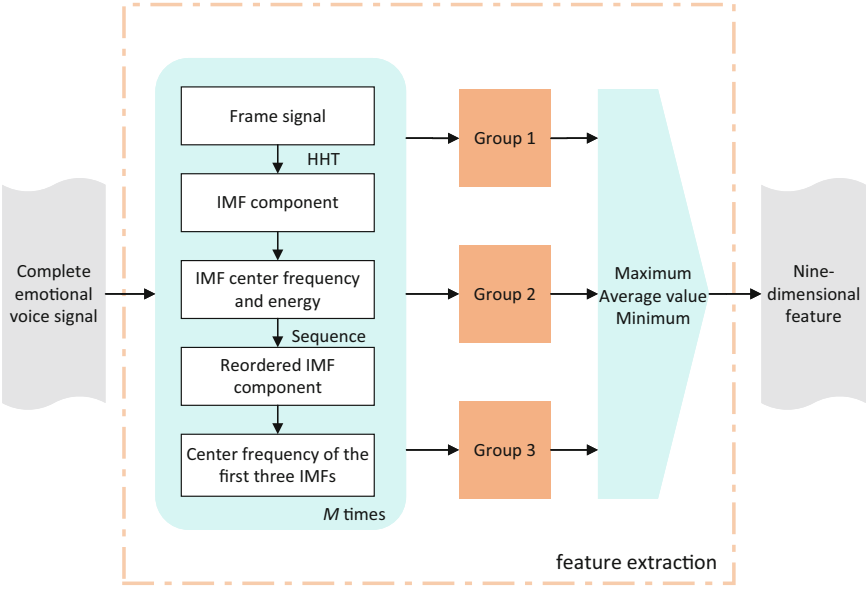

Fig. 1. Flow chart of emotion feature extraction.

3.2 Multi-dimensional Feature Extraction

We believe that the energy, as well as the magnitude and distribution of the central frequencies in the speech signal, are of great importance for distinguishing different emotions. Therefore, to extract the relevant features, we performe HHT transform on each frame signal and rearrange the IMF components according to the energy intensity from largest to smallest to extract their center frequencies, naming them as first center frequency, second center frequency and third center frequency in turn. Due to the complexity of emotional features in speech signals, each individual frame signal cannot fully express the features of the entire sentence. Therefore, the three central frequencies of all frames in the sentence are divided into three groups, and three data points for maximum, minimum, and average values are extracted from each group. The three sets of nine data are used as sentence features. This is the simplest method to extract overall features from all frame features. As shown in Fig. 1, *Groups* 1, 2, and 3 are the three central frequency grouping cases of all frame signals in a speech sentence. If complete speech is split into M-frame signals, there are M center frequencies in *Groups* 1, 2, and 3.

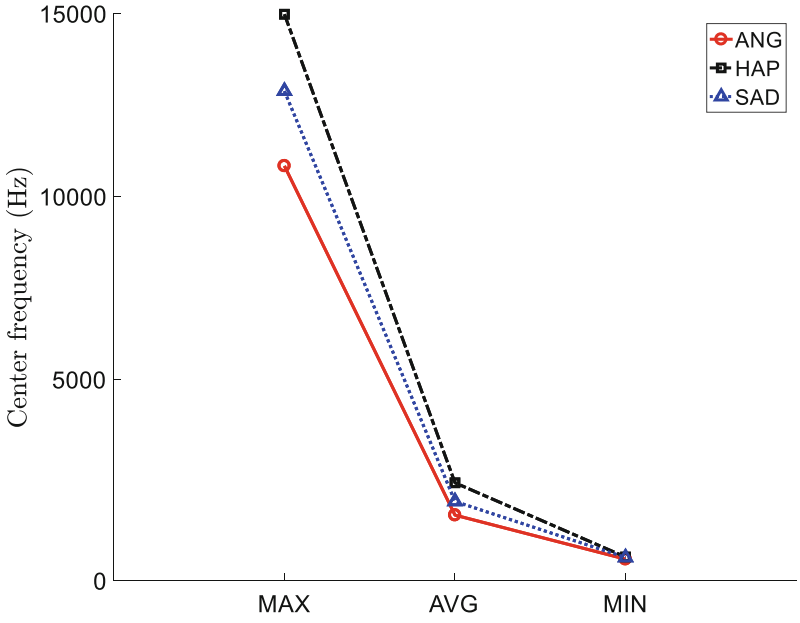

Fig. 2. Comparison of values in *Group* 1 for the three emotions.

The voice signal in the data pool is essentially devoid of noise interference, so it is possible to directly perform HHT transform on each IMF component.

$$\widehat{IMF_i(t)} = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{IMF_i(\tau)}{t-\tau} \mathrm{d}\tau \tag{9}$$

The complex resolved signal for the i-th IMF component can be constructed from $IMF_i(t)$ and $\widehat{IMF_i(t)}$ as

$$z_i(t) = IMF_i(t) + \mathrm{i}\widehat{IMF_i(t)} = a_i(t)\mathrm{e}^{\mathrm{i}\theta_i(t)} \tag{10}$$

$$a_i(t) = \sqrt{IMF_i(t)^2 + \widehat{IMF_i(t)}^2} \tag{11}$$

$$\theta_i(t) = \arctan \frac{\widehat{IMF_i(t)}}{IMF_i(t)} \tag{12}$$

In the above equation, $a_i(t)$ is the instantaneous amplitude and $\theta_i(t)$ is the instantaneous phase, from which the instantaneous frequency is obtained as follows:

$$f_i(t) = \frac{1}{2\pi} \cdot \frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t} \tag{13}$$

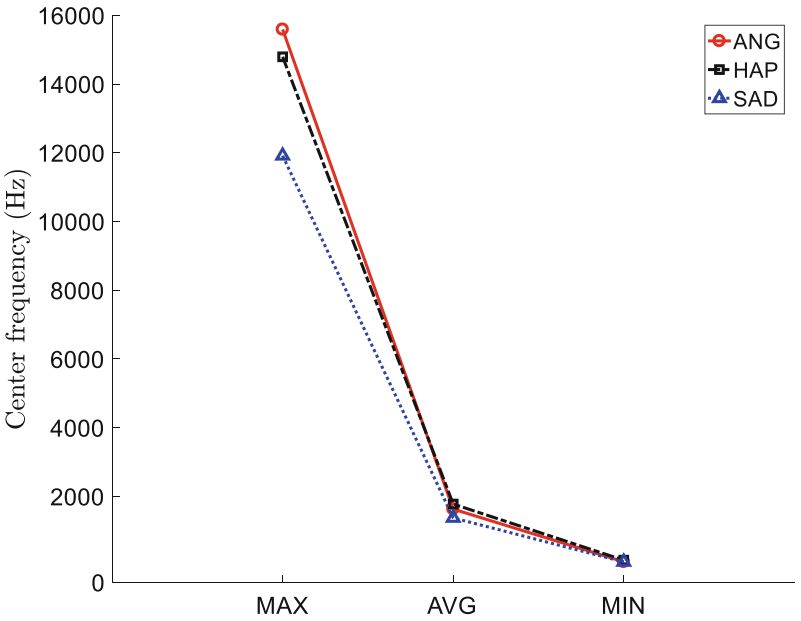

Fig. 3. Comparison of values in *Group* 2 for the three emotions.

If the IMF component has N sampling points, then let the instantaneous frequency at each sampling point be f_{in} and the instantaneous amplitude be a_{in}, then the instantaneous energy at each sampling point and the average energy of the whole IMF can then be obtained as follows:

$$E_{in} = {a_{in}}^2 \tag{14}$$

$$\overline{E_i} = \frac{\sum_{n=1}^{N} E_{in}}{N} \tag{15}$$

Then, after sorting each IMF component by average energy from largest to smallest, the central frequency of each IMF can be defined by the weighted value of the instantaneous frequency.

$$\overline{f_i} = \frac{\sum_{n=1}^{N} (E_{in} f_{in})}{\sum_{n=1}^{N} E_{in}} \tag{16}$$

The center frequencies of the top three of these IMF components were selected as the first, second, and third center frequency.

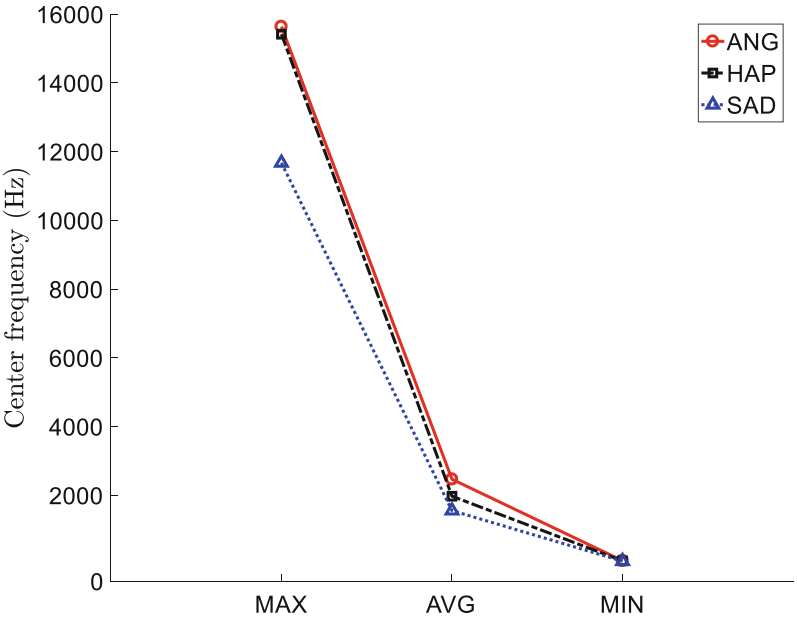

Fig. 4. Comparison of values in *Group* 3 for the three emotions.

The three center frequencies of all frames of speech are divided into three groups according to the above requirements, and the maximum f_{max}, minimum f_{min} and average f_{avg} of each group are extracted.

$$f_{max} = \max(f_1, f_2, ..., f_M) \tag{17}$$

$$f_{min} = \min(f_1, f_2, ..., f_M) \tag{18}$$

$$f_{avg} = \text{average}(f_1, f_2, ..., f_M) \tag{19}$$

f_1,f_2,...,f_M in the above equation are the center frequencies of the M frame signals, and a total of nine data points from the three groups are combined as the

sentiment feature parameters of speech. Finally, the nine-dimensional sentiment feature parameters of all speech are input into the SVM classification algorithm model for training and testing.

Figures 2, 3, and 4 show the maximum, mean, and minimum values extracted from *Groups* 1, 2, and 3. Figure 2 shows that the maximum and average values of happy emotion are the highest, and angry emotion was the lowest, while the minimum values of the three emotions are similar. Figure 3 shows that the maximum of angry emotion is the lowest and the average value of happy emotion is the highest, while the minimum values of the three emotions are similar. Figure 4 shows that the maximum value of happy emotion is the highest, the average of angry emotion was higher than of happy, and the average of happy emotion is higher than of sad. It is clear that the above nine characteristic values of different emotions have different distributions, we can identify emotions based on the above three groups of nine data points in total.

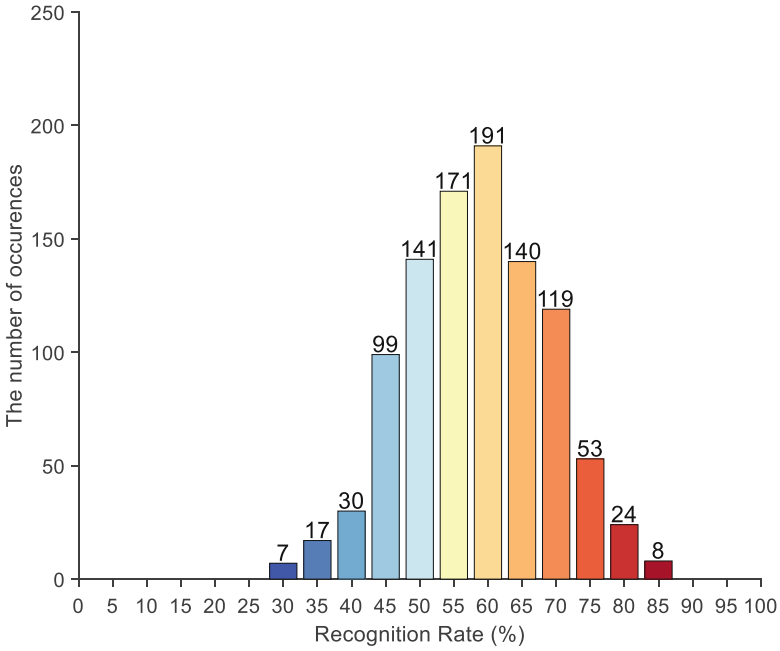

Fig. 5. Distribution of recognition rate of 1000 KNN random tests after 15 ms frame length with 25% frame shift and frame splitting processing.

4 Classification Algorithm

To further demonstrate the effectiveness of the ICEEMDAN and multi-IMF centerfrequency-based feature extraction methods and to investigate the effect

of frame splitting on feature extraction, we utilized two classification techniques: the k-nearest neighbor algorithm (KNN) and the support vector machine algorithm (SVM) [29,30].

At the heart of the KNN algorithm is the concept of using the labels of the k-nearest neighbors from the training set to a test sample as the basis for the prediction of the label of the test sample. By employing a kernel function, the SVM algorithm maps the sample points of input features onto a high dimensional feature space, where all training samples are positioned above or below a hyperplane that maximizes the margin between different categories; the data are partitioned by the hyperplane to achieve classification. Each of these two classification algorithms has its own advantages and disadvantages, but they are suitable for small databases. Therefore, we try these two algorithms to process the data.

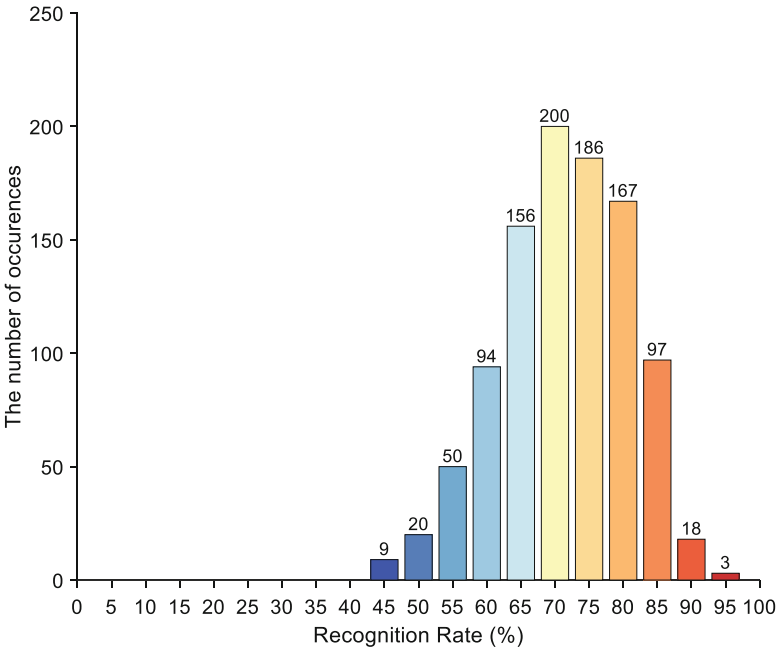

Fig. 6. Distribution of recognition rate of 1000 SVM random tests after 15 ms frame length with 25% frame shift and frame splitting processing.

Most traditional frame lengths for speech signals range from 10 to 30 ms, without a uniform standard. First, we use a 15 ms frame length with a 25% frame shift to process the signal and analyzed the results. For the KNN classification experiment, we randomly test 1000 times using 80% of the samples for training and 20% for testing. We ensure that the three emotions are evenly distributed in the training samples to improve the training effectiveness. The results

are shown in Fig. 5, indicating that the emotion recognition rate is mainly distributed around 60%, but lower recognition rates also appeared several times.

Table 1. Recognition rate of actor CC using SVM algorithm for five frame lengths.

Frame length (ms)	Frame Shift (%)	Recognition rate (%)
15	25	71.4
30	25	75.6
45	25	71.0
60	25	62.4
75	25	56.6

Subsequently, we used the SVM algorithm to perform 1000 classification recognition tests on the characteristic parameters of the same sample data; Fig. 6 shows the statistical results. The recognition rates of the main occurrences are 70% and 75%, and they appeare 200 and 186 times, respectively, which basically occupies half of the 1000 tests. Based on a comparison of the two algorithms, we find that the KNN algorithm is not suitable for our sample data. Therefore, we proceed with the SVM algorithm for the feature parameters in subsequent framing studies.

Table 2. Recognition rate of actor CC using SVM algorithm for three frame shifts.

Frame length (ms)	Frame Shift (%)	Recognition rate (%)
30	25	75.6
30	50	72.8
30	75	68.6

In Table 1, we compare the average recognition rate of 1000 SVM tests with different framing methods of the actor CC and find that the recognition rate in the case of framing with a 30 ms frame length with a 25% frame shift reached 75.6%, which is significantly higher than of other framing methods. Figure 7 shows the recognition rate statistics of the 1000 tests with a 30 ms frame length 25% frame shift, with the main recognition rates appearing being 75% and 80%, and they appear 204 and 185 times, respectively, occupying half of the number of 1000 tests. Comparing to Fig. 6, it can be clearly seen that a 30 ms frame length with a 25% frame shift is more effective.

In Table 2, we compare the recognition rates for different frame shifts for a frame length of 30 ms. It is found that the recognition rate decreases as the frame shift increases, which represents the continuity of the features of the emotional speech signal; therefore, it is necessary to smooth over each frame signal when

intercepting the signal. In addition, we do the same for the speech data of another actor GRW in the database and investigated the effect of different frame lengths and frame shifts on the recognition rate. The recognition rates obtained using 1000 SVMs for the six frame-splitting methods are presented in Table 3. It can be seen that the recognition rate reaches 65.4% at a frame length of 30 ms and frame shift of 25%, which is still the highest recognition rate among all the frame-sharing methods. The overall recognition rate of the actor GRW speech data is not as good as that of the actor CC speech data, but the frame length and frame shift still have an impact on the recognition rate.

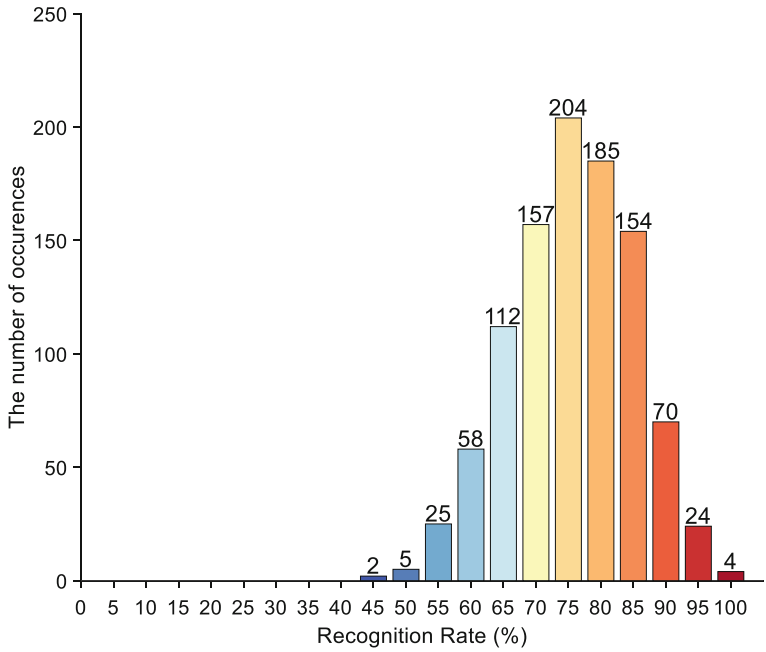

Fig. 7. Distribution of recognition rate of 1000 SVM random tests after 30 ms frame length with 25% frame shift and frame splitting processing.

The classification outcomes of our dataset indicate that our feature extraction method is effective in recognizing emotional speech signals. Moreover, the implementation of a framing method utilizing a 30 ms frame length with a 25% frame shift proves to be beneficial in improving the recognition accuracy of emotional speech signals during signal processing.

Table 3. Recognition rate of actor GRW using SVM algorithm with different frame length and frame shift.

Frame length (ms)	Frame Shift (%)	Recognition rate (%)
15	25	64.5
30	25	65.4
30	50	62.2
30	75	61.3
45	25	59.4
60	25	58.6
75	25	56.3

5 Conclusion

In order to effectively study the impact of framing on feature extraction, we propose a novel combined recognition method that utilizes the ICEEMDAN, multi-IMF CFP feature extraction, and SVM algorithm. Our approach employs the ICEEMDAN algorithm to overcome the mode mixing issue associated with traditional EMD techniques, reduces reconstruction errors of EEMD, and eliminates false modes in CEEMDAN, thus improving the accuracy and efficiency of signal decomposition. Additionally, our combination of center frequency and IMF ranking offers an effective means of analyzing emotional speech signals, while also enhancing the completeness and integrity of decomposed IMFs. Simultaneously considering the local features and global features of speech signals is a key step in analyzing the emotional features in speech signals. We recombine the features of all frame signals of complete speech into a new set of features, which can incorporate local features without losing global features. Ultimately, our emotional speech signal classifier, constructed using the SVM algorithm, achieved an overall recognition accuracy of 75.6% during the recognition of 1000 random tests.

Despite successfully identifying emotional speech signals within the constructed data pool, the range of emotions analyzed is limited, and the recognition rate falls short of current state-of-the-art emotional speech recognition methods. The proposed method uses improved EMD to decompose speech samples into IMFs, rearranges them based on energy intensity, calculates central frequencies, and extracts multidimensional features. These features are then classified using a SVM, thus achieving an effective analysis of the emotional features of speech signals, and these steps can be used to identify more emotional types. The algorithm can be used without any prior knowledge for the classification of emotional speech signals and can identify emotion types based on the classification results. We have also study the effects of five frame lengths and three frame shifts on the recognition rate according to this feature extraction method. The best framing method obtained had a frame size of 30 ms and a frame shift of 25%.

References

1. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Li, H., Ching, P. (eds.) INTERSPEECH 2014, Interspeech, vol. 1-4, pp. 223-227. ISCA, BAIXAS (2014)
2. Li, R., Wu, Z., Jia, J., Zhao, S., Meng, H.: Dilated residual network with multi-head self-attention for speech emotion recognition. In: ICASSP 2019, pp. 6675–6679. IEEE, NEW YORK (2019)
3. Satt, A., Rozenberg, S., Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms. In: INTERSPEECH 2017, Interspeech, vol. 1-6, pp. 1089-1093. ISCA, BAIXAS (2017). https://doi.org/10.21437/Interspeech.2017-200
4. Zhong, Y., Hu, Y., Huang, H., Silamu, W.: A lightweight model based on separable convolution for speech emotion recognition. In: INTERSPEECH 2020, Interspeech, vol. 11, pp. 3331-3335. ISCA, BAIXAS (2020). https://doi.org/10.21437/Interspeech.2020-2408
5. Akcay, M.B., Oguz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020)
6. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP 2013, pp. 6643-6649. IEEE, NEW YORK (2013)
7. Huang, K.Y., Wu, C.H., Hong, Q.B., Su, M.H., Chen, Y.H.: Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: ICASSP 2019, pp. 5866-5870. IEEE, NEW YORK (2019)
8. Abdullah, S.M.S.A., Ameen, S.Y.A., Sadeeq, M.A., Zeebaree, S.: Multimodal emotion recognition using deep learning. J. Appl. Sci. Technol. Trends **2**(02), 52–58 (2021)
9. Alnuaim, A.A., Zakariah, M., Shukla, P.K., Alhadlaq, A., Hatamleh, W.A., Tarazi, H., Sureshbabu, R., Ratna, R.: Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. J. Healthcare Eng. **2022**, 6005446 (2022)
10. Nayak, S., Nagesh, B., Routray, A., Sarma, M.: A human-computer interaction framework for emotion recognition through time-series thermal video sequences. Comput. Electr. Eng. **93**, 107280 (2021)
11. Chowdary, M.K., Nguyen, T.N., Hemanth, D.J.: Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications pp. 1–18 (2021)
12. Chen, H., Zhang, B.: Adaptive algorithm for feature selection of speech emotion recognition based on genetic algorithm and svm. J. Phys. Conf. Ser. **1883**(1), 012019 (2021)
13. Korkmaz, O.E., Atasoy, A.: Emotion recognition from speech signal using mel-frequency cepstral coefficients. In: 9th international conference on electrical and electronics engineering (ELECO), pp. 1254-1257. IEEE, NEW YORK (2015)
14. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. Math. Phys. Eng. Sci. **454**(1971), 903–995 (1998)
15. Tamulevičius, G., Korvel, G., Yayak, A.B., Treigys, P., Bernatavičienė, J., Kostek, B.: A study of cross-linguistic speech emotion recognition based on 2d feature spaces. Electronics **9**(10), 1725 (2020)

16. Wu, Z., Huang, N.E.: A study of the characteristics of white noise using the empirical mode decomposition method. Proc. Math. Phys. Eng. Sci. **460**(2046), 1597–1611 (2004)
17. Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M.A., Cleder, C.: Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. Speech Commun. **114**, 22–35 (2019)
18. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. **1**(1), 1–41 (2009)
19. Chen, J., Li, H., Ma, L., Bo, H., Gao, X.: Application of eemd-hht method on eeg analysis for speech evoked emotion recognition. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 376-381. IEEE, Shenzhen (2020)
20. Yeh, J.R., Shieh, J.S., Huang, N.E.: Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. Adv. Adapt. Data Anal. **2**(2), 135–1122 (2010)
21. Han, T., Liu, Q., Zhang, L., Tan, A.C.: Fault feature extraction of low speed roller bearing based on teager energy operator and ceemd. Measurement **138**, 400–408 (2019)
22. Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P.: A complete ensemble empirical mode decomposition with adaptive noise. In: ICASSP 2011, pp. 4144-4147. IEEE (2011)
23. Gao, B., Huang, X., Shi, J., Tai, Y., Zhang, J.: Hourly forecasting of solar irradiance based on ceemdan and multi-strategy cnn-lstm neural networks. Renew. Energy **162**, 1665–1683 (2020)
24. Colominas, M.A., Schlotthauer, G., Torres, M.E.: Improved complete ensemble emd: a suitable tool for biomedical signal processing. Biomed. Signal Process. Control **14**, 19–29 (2014)
25. Alimuradov, A.K., Tychkov, A.Y., Makarova, N.A.: Study of voiced speech using empirical mode decomposition to detect stressful emotions in human-robot interaction. In: 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), pp. 7-10. IEEE (2020)
26. Jin, S.Y., Su, Y., Guo, C.J., Fan, Y.X., Tao, Z.Y.: Offshore ship recognition based on center frequency projection of improved emd and knn algorithm. Mech. Syst. Signal Process. **189**, 110076 (2023)
27. Liu, P., Pell, M.D.: Recognizing vocal emotions in mandarin chinese: a validated database of chinese vocal emotional stimuli. Behav. Res. Methods **44**, 1042–1051 (2012)
28. Shome, N., Barlaskar, S.A., Laskar, R.H.: Significance of frame size and frame shift on vowel on set point detection. In: IEEE International Conference on Recent Trends in Electronics, pp. 1272-1276. IEEE (2016)
29. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R.K., et al.: Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590 (2020)
30. Umamaheswari, J., Akila, A.: An enhanced human speech emotion recognition using hybrid of prnn and knn. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 177-183. IEEE (2019)