



# Long-Time Speech Emotion Recognition Using Feature Compensation and Accentuation-Based Fusion

Jiu Sun<sup>1</sup> · Jinxin Zhu<sup>1</sup> · Jun Shao<sup>1</sup>

Received: 2 June 2023 / Revised: 26 July 2023 / Accepted: 27 July 2023 /

Published online: 11 September 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

In this paper, we study the speech emotion feature optimization using stochastic optimization algorithms, and feature compensation using deep neural networks. We also proposed to use accentuation-based fusion for long-time speech emotion recognition. Firstly, the extraction method of emotional features is studied, and a series of speech features are constructed for the recognition of emotion. Secondly, we propose a method of sample adaptation through denoising autoencoder to enhance the versatility of features through the mapping of sample features to improve adaptive ability. Thirdly, GA and SFLA are used to optimize the combination of features to improve the emotion recognition results at the utterance level. Finally, we use transformer model to implement accentuation-based emotion fusion in long-time speech. The continuous long-time speech corpus, as well as the public available EMO-DB, are used for experiments. Results show that the proposed method can effectively improve the performance of long-time speech emotion recognition.

**Keywords** Speech emotion recognition · Feature compensation · Long-time emotion recognition · Accentuation-based fusion

## 1 Introduction

Speech emotion recognition represents a critical area of research. Emotions serve as essential elements in human communication and expression, exerting considerable

---

✉ Jiu Sun  
sunjiu@ycit.edu.cn

Jinxin Zhu  
zjx@ycit.edu.cn

Jun Shao  
sj@ycit.edu.cn

<sup>1</sup> School of Information Technology, Yancheng Institute of Technology, Yancheng 224051, China

influence on an individual's behavior and psychological well-being [18]. Understanding and accurately identifying emotions conveyed through speech are therefore of great significance. Emotion recognition has a wide range of applications in many fields, such as human–computer interaction, diagnosis and treatment of mental illness, social media analysis, and more.

Long-time speech emotion recognition studies the temporal changes of emotions over a paragraph period of time. This is a highly challenging problem, as traditional algorithms mainly focus on the emotional state within a single sentence, losing the emotional information over time in the context [10].

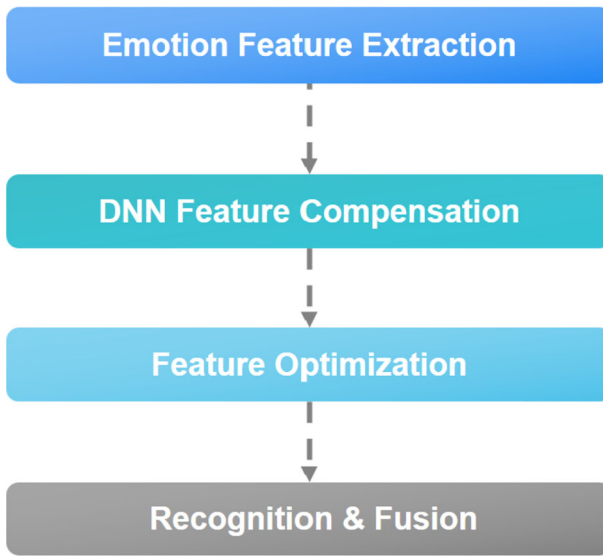
The extraction, compensation, and optimization of speech emotion features constitute pivotal challenges in the field of speech emotion recognition. Addressing these challenges necessitates a holistic approach integrating signal processing and machine learning techniques. Speech emotion feature extraction involves the derivation of features from speech signals that effectively capture emotional characteristics. Key speech emotion features commonly utilized include pitch, formant, speaking rate, energy, and intonation [6, 17].

Feature extraction based on Mel frequency cepstral coefficients (MFCC) is a method that converts speech signals into Mel frequency spectrograms and extracts spectral coefficients as features. The method based on intonation analysis is a feature extraction method that extracts emotion features by analyzing the changes in pitch of speech signals. Short-term energy analysis is a feature extraction method that extracts emotion features by analyzing the changes in acoustic energy of speech signals. Using speech duration, we can construct feature extraction method that extracts emotion features by analyzing the duration of different syllables in speech signals.

Apart from feature extraction, compensation and optimization of speech emotion features are also important. Especially in practical applications, due to differences in speech features among different individuals, it is necessary to compensate and optimize these differences. There has been relatively less research on the compensation of emotional features, and previous research mainly focused on normalization of features to compensate for the impact of individual differences [23].

The optimization of features can play a crucial role in enhancing the quality of speech emotion features and significantly improving the accuracy of speech emotion recognition systems. Common feature selection methods include correlation coefficient, mutual information, and so on. In addition, random optimization methods such as genetic algorithm and swarm intelligence have important value for selecting emotional features [19]. Such algorithms randomly select feature combinations from a feature subset to obtain a new feature subset. If the performance of the new feature subset is better than the current solution, the new solution is accepted. This step is repeated until a specified stopping criterion is reached. Random optimization methods can search for the optimal solution by repeatedly sampling, evaluating, and updating the solution. It can effectively solve the feature selection problem, especially for searching and optimizing cases with a large number of emotional features.

Researchers have investigated the practical applications of emotion recognition and have conducted comparative analyses of various modeling algorithms in this domain. Albu et al. [3] explore various neural network approaches for children's emotion recognition, specifically focusing on speech signals and facial images. It highlights



**Fig. 1** Long-time speech emotion recognition

the influence of the number of centers chosen by the k-means algorithm on the recognition performance of radial basis function (RBF) networks and extreme learning machines (ELM). The findings emphasize the importance of child affective modeling alongside cognitive modeling for intelligent software applications and technology-enhanced learning, indicating the implications for personalized tutoring.

In the context of long-time speech, the challenge in emotion recognition addressed by the paper lies in optimizing speech emotion feature extraction and compensation. This involves dealing with complex emotions expressed in long-time speech and capturing the dynamic relationships between emotions over time.

In this paper, we study the extraction, compensation, and optimization of emotional features in speech and their application in long-time speech emotion recognition. In order to improve emotion recognition in continuous long-time speech utterances, we studied a novel framework involving four main steps, as shown in Fig. 1: emotional feature extraction, sample adaptation using neural network, feature optimization using GA and SFLA, and a novel accentuation-based fusion method for long-time speech emotion recognition.

In the first step, a series of speech features are extracted for emotion recognition at the frame level. These features include prosodic and spectral features, such as pitch, intensity, formant, and MFCCs.

In the second step, a neural network is used for sample adaptation to enhance the versatility of features. The neural network maps the sample features to a set of latent variables that capture the underlying emotional content of the speech. The features of a new speech sample are adapted to improve adaptive ability.

In the third step, GA and SFLA are used to optimize the combination of features to improve the accuracy of emotion recognition. GA and SFLA are metaheuristic

optimization algorithms that can search for the optimal solution in a large solution space.

In the fourth step, we propose to use a novel accentuation-based fusion algorithm to combine context information and accentuation information in long-time speech emotion. Each utterance is modeled independently and jointly recognized for the final emotion category.

We introduce a unique method that represents emotions as nodes and transitions as edges in a graph, utilizing the transformer model's self-attention mechanism to predict future states based on previous states and encoder outputs. The incorporation of accentuation weights enhances emotion recognition accuracy, offering a promising solution for understanding and recognizing emotions in extended speech contexts.

The proposed method was tested on continuous long-time speech utterances, and the results showed that it effectively improved the accuracy of emotion recognition. By optimizing the combination of features and enhancing the versatility of features through sample adaptation, the proposed method was able to improve the performance of emotion recognition in continuous long-time speech utterances.

Overall, we studied a comprehensive approach to improving emotion recognition in continuous long-time speech utterances by addressing the challenges of feature compensation, optimization, and accentuation-based results fusion. The method can be applied in a variety of contexts, such as emotion recognition in therapy, education, and customer service.

## 1.1 Related Work

Many existing emotion recognition studies have considered the problem of feature selection and feature analysis [2, 14, 15, 21, 25]. Alex et al. [4] studied feature selection in utterance level and syllable level for emotion recognition. Abdelhamid et al. [1] studied stochastic optimization for speech emotion models. Zhang et al. [28] proposed to study the practical speech emotion using stochastic optimization algorithms. In their study, basic feature was analyzed and feature combination was used to model speech emotions. They further studied emotion types that had practical values. Xu et al. [26] studied a large set of speech emotional features, and the results were promising. Although the novelty in the graph learning-based classifier was high, the generalization ability of the algorithms needed to be further discussed. Gat et al. [11] studied speaker feature normalization for emotion models. Huang et al. [13] studied feature normalization using speaker-sensitive features. A general framework was proposed to improve the emotion recognition performance. Saad et al. [24] studied emotion recognition across different languages and databases. The transfer of models is a very interesting topic. The optimal set of emotional features need to be further studied. Cowen et al. [8] studied a large number of emotions with feature analysis. Hajarolasvadi et al. [12] studied convolution neural network and its application in spectrograms. They propose to model emotions using visual features of the spectrogram. Fahad et al. [9] studied speaker-adaptive SER system. They presented a promising solution to the issue of speaker variability, enhancing accuracy in emotion recognition tasks.

Feature space maximum likelihood linear regression was used, and emotion-specific epoch-based features are explored.

Other researchers have focused on the emotion models. Zou et al. [29] studied the cognitive-related emotions and the detection from speech. They propose to record the oral report during math exercises and analyzed the emotional features. Although the results were promising, the relation between acoustic features and the cognitive states still needs investigation. Anvarjon et al. [5] studied a lightweight detector based on a novel CNN architecture. Although the results were promising, more variety of backbone networks could be discussed with novel emotional features. Jin et al. [16] studied support vector machines from a semi-supervised framework. They propose to apply self-training SVM to speech emotion recognition and tested on public available databases. Although the results were promising, more recent algorithms needed to be further discussed. Choudhary et al. [7] studied emotion recognition using deep neural networks. The representative learning requires a large number of training samples. Although the results were promising, the generalization of the model is dependent on the dataset. Oaten et al. [20] studied a special type of emotion, disgust and its practical values in health.

## 2 Methodology

### 2.1 Feature Compensation

The problem of uneven sample distribution is an important challenge in feature engineering. In the process of emotion modeling, we often need a large number of samples, so that the statistical distribution we learn is very consistent with the real situation. However, our sample data sets are often limited, and the distribution of samples is uneven from different angles, such as age, accent, and personality. This will directly lead to the model we learned, which is not highly versatile, and the effect on the new sample is difficult to guarantee. Therefore, it is necessary to study the compensation method of the features to compensate for the equilibrium problem of the sample.

Deep neural networks are used to normalize and compensate for features, so that the imbalance distribution in the sample is alleviated, as shown in Fig. 2. Feature-compensated samples can be better counted and modeled. The input of the network is a one-dimensional emotional feature vector of each sample, and the output of the network is expressed by supervised information, which is also the emotional feature vector after one-dimensional compensation.

Feature compensation algorithms play a particularly important role. In the noise scenario, we add various types of noise to the test sample, which destroys the original emotional features to a certain extent. The training samples were collected in a relatively quiet environment, and the signal was relatively pure. The problem of sample mismatch caused by noise is a great application bottleneck in the practical application of speech emotion recognition. The method of deep network feature compensation can solve the feature mapping from pure speech to noisy speech, or vice versa. To a certain extent, the influence of noise is reduced. It should be pointed out that the current emotional database rarely exists in an absolutely quiet environment, inevitably,

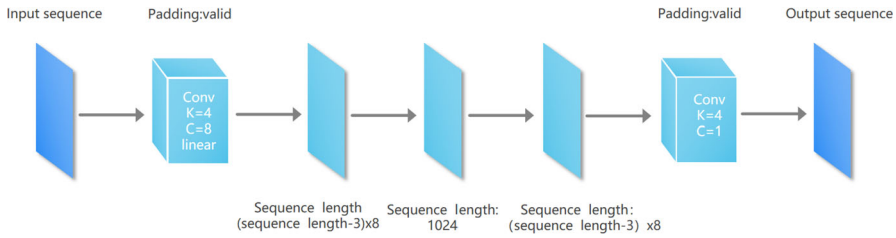


Fig. 2 Denoising autoencoder for feature compensation

the training data center brings all kinds of noise, so in practical applications, the test environment can be of higher quality than the training data, such as on a smartphone, close collection of voice can obtain high quality. Therefore, the mismatch of noise is not necessarily unidirectional, the sound quality of the test environment can also be better than the training corpus, and simple noise reduction methods cannot replace the feature compensation algorithm.

### 2.2 Feature Optimization

When it comes to the selection and combination of features, it is difficult to exhaust the full range of possibilities. Therefore, in this paper, stochastic optimization algorithms are used to optimize the sample features. We compare the GA and SFLA algorithms and optimize the combination of features to improve the emotion recognition results at the frame level.

#### GA-Based Feature Selection

The fitness function is defined in Eq. 1:

$$fitness_i = f(\mathbf{features}_i) \tag{1}$$

where  $fitness_i$  is the fitness of the  $i$ -th individual,  $\mathbf{features}_i$  is the binary vector representing the presence or absence of each feature for the  $i$ -th individual, and  $f()$  is the performance metric of the model trained on the selected features.

The selection of individuals for crossover and mutation using Roulette wheel selection:  $p_i = \frac{fitness_i}{\sum_{j=1}^{popsize} fitness_j}$ , where  $p_i$  is the probability of selecting the  $i$ -th individual,  $fitness_i$  is the fitness of the  $i$ -th individual, and  $\sum_{j=1}^{popsize} fitness_j$  is the sum of fitness values in the population.

The crossover of two individuals using a one-point crossover operator, as shown in Eq. 2, is:

$$\mathbf{c}_1, \mathbf{c}_2 = \text{crossover}(\mathbf{p}_1, \mathbf{p}_2, p_c) \tag{2}$$

where  $\mathbf{p}_1$  and  $\mathbf{p}_2$  are the parent binary vectors,  $\mathbf{c}_1$  and  $\mathbf{c}_2$  are the offspring binary vectors, and  $p_c$  is the crossover probability.

The mutation of an individual using a bitwise mutation operator is shown in Eq. 3.

$$\mathbf{m} = \text{mutation}(\mathbf{p}, p_m) \tag{3}$$

where  $\mathbf{p}$  is the parent binary vector,  $\mathbf{m}$  is the mutated binary vector, and  $p_m$  is the mutation probability.

The replacement of the worst individuals in the population with the best individuals from the new population, Eq. 4, is:

$$\begin{aligned} &\text{if } fitness_{\text{new}} > fitness_{\text{worst}} \\ &\text{then } \text{replace}(worst_{\text{individual}}, new_{\text{individual}}) \end{aligned} \quad (4)$$

where  $fitness_{\text{new}}$  is the fitness of the new individual,  $fitness_{\text{worst}}$  is the fitness of the worst individual in the population, and  $\text{replace}()$  is a function that replaces the worst individual with the new individual.

### SFLA-Based Feature Selection

The fitness function is defined the same as the one used for GA (Eq. 1), in which  $fitness_i$  is now the fitness of the  $i$ -th frog, and  $\mathbf{features}_i$  is the binary vector representing the presence or absence of each feature for the  $i$ -th frog.

The creation of a new frog population by combining the best two frogs from each memplex is shown in Eq. 5.

$$\mathbf{x}_{\text{new}} = (\mathbf{x}_{\text{best1}} \& \mathbf{x}_{\text{best2}}) | \sim (\mathbf{x}_{\text{best1}} | \mathbf{x}_{\text{best2}}) \quad (5)$$

where  $\mathbf{x}_{\text{best1}}$  and  $\mathbf{x}_{\text{best2}}$  are the binary vectors of the best two frogs in the memplex,  $\&$  is the bitwise AND operator,  $|$  is the bitwise OR operator, and  $\sim$  is the bitwise NOT operator.

The replacement of the worst frogs in each memplex with the best frogs from the new population is shown in Eq. 6:

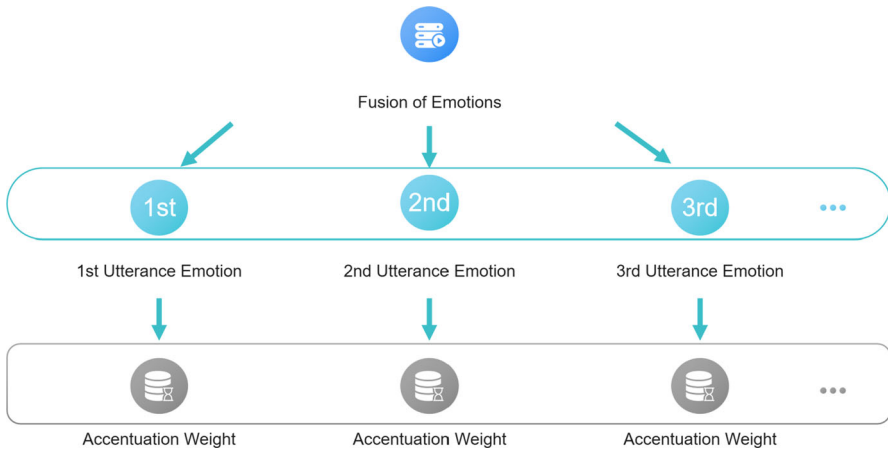
$$\begin{aligned} &\text{if } fitness_{\text{new}} > fitness_{\text{worst}} \\ &\text{then } \text{replace}(worst_{\text{frog}}, new_{\text{frog}}) \end{aligned} \quad (6)$$

where  $fitness_{\text{new}}$  is the fitness of the new frog,  $fitness_{\text{worst}}$  is the fitness of the worst frog in the memplex, and  $\text{replace}()$  is a function that replaces the worst frog with the new frog.

## 2.3 Accentuation-Based Fusion of Emotions

In order to better analyze the utterance-level characteristic of emotions, we use a graphical chain model to represent emotional behavior in long-time speech. We can capture the complex and dynamic relationships between different aspects of emotions and build more accurate models for long-time emotion recognition.

In continuous speech, emotion states also change continuously. At the conventional frame-level recognition, emotion labels are assigned to short-time periods. However, in long-time periods of utterances, these emotion recognition results should be merged. Previous studies on continuous speech have been focused on the linguistic meaning [22]; the parallel-linguistic information, such as emotion, is not well studied. Since emotions in speech typically last around one to several seconds, fusion of neighboring



**Fig. 3** Emotion labels fusion with accentuation weights

emotion labels is a direct implementation of long-time speech emotion recognition, as shown in Fig. 3.

For example, each node in the graph could represent a specific emotion label such as happy, sad, angry, and neutral, which is assigned to a specific utterance.

$$\text{EmotionSequence} = \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \dots, \mathbf{e}_m \quad (7)$$

where  $\mathbf{e}_i$  represents an emotional label of an utterance.

The edges between these nodes could represent the transitions between these emotions. The weights on the edges could represent the likelihood of transitioning from one emotion to another, based on the emotional behavior characteristics of the speech.

Using a graphical chain model, we can estimate the transition probability. By estimating probability of the next emotion labels, conditioned on the previous observation of label sequence, we can build a predictor to identify the change of emotions in long-time speech. Errors in segment-level emotion recognition can be corrected, when abnormal edges with low posterior probability are detected.

Transformer model is adopted to build the predictor. The transformer is a type of neural network architecture that can be used for various sequential data processing. It is a feedforward neural network that uses self-attention to process inputs (emotion label sequence in long-time speech emotion).

The attention mechanism works by assigning a weight to each element of the input sequence based on its relevance to the output at each step of the processing. The weights are calculated using a function that takes into account the similarity between the current processing step and each element of the input sequence.

The prediction algorithm steps are as follows:

**Input:**

- Bi-direction Emotion label sequence;
- Accentuation Weights Sequence.



The accentuation weights are related to each emotion label, and they can be calculated by utterance-level acoustic features.

#### **Output:**

Predict the transition probability of the next node;

If it is lower than an empirical threshold, replace it with the predicted emotion label.

To predict the next state of emotion sequence in long-time speech, we can first input the sequence into the transformer encoder to obtain a sequence of encoder outputs. Then, we can use the decoder to generate the next state based on the previous states and the encoder outputs. Specifically, at each time step, the decoder generates an output representation based on the previous output and the encoder outputs and then generates a probability distribution over the possible next states using a softmax function.

To estimate the probability of state transitions, we can compute the probability of transitioning from the current state to each possible next state using the output distribution generated by the decoder. The transition probabilities can then be used to construct a state transition matrix that describes the probability of transitioning between any two states in the sequence.

Furthermore, we consider the accentuation, which is a cue of important utterance in a paragraph.

To identify long-period emotion type over a paragraph, we consider the accentuation weights. A sliding window is used to generate the samples. The weights are estimated by a regression model, which reflects the accentuation features. The accentuation-related features used for regression include pitch frequency, formant frequency, duration time, and intensity.

By using a graph to represent emotions in this way, we can capture the dynamic changes in emotional expression over time and build models that can improve the recognition of emotional state of the speaker for long period of speech.

## **3 Experimental Results**

### **3.1 The Databases**

In our experiment, we use EMO-DB for feature compensation experiment and emotion recognition test.

EMO-DB is a widely used emotional speech database that contains recordings of emotional speech in German. It was created by the Institute of Communication Science and Phonetics at the University of Bonn in Germany.

The EMO-DB database comprises seven emotions: (1) anger; (2) boredom; (3) disgust; (4) fear; (5) happiness; (6) sadness; and (7) neutral. The data were recorded at a 16-kHz sampling rate.

EMO-DB has been used in various studies to analyze emotional speech and develop algorithms for speech emotion recognition. It is freely available for academic research purposes and has been used in numerous studies worldwide.

We adopt another local database from Southeast University, whose long-time speech corpus is an idea to verify our emotion recognition method.

**Table 1** The original sample distribution of SEU database

Emotion types	Gender	Short sentence	Long paragraph
Happy	Male	238	18
	Female	242	18
Fear	Male	220	18
	Female	234	18
Sad	Male	230	18
	Female	232	18
Surprise	Male	239	15
	Female	241	15
Angry	Male	240	18
	Female	237	18
Neutral	Male	243	18
	Female	243	18

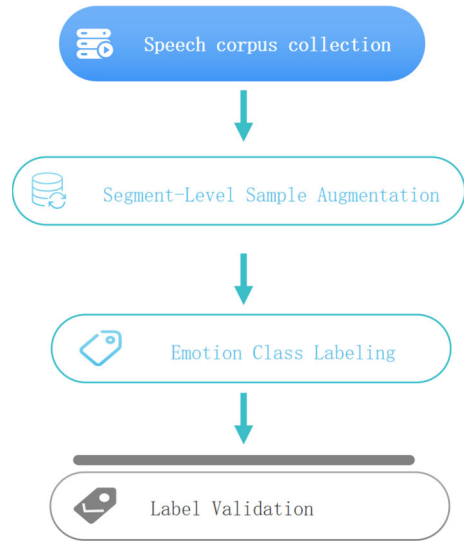
The database involves three males and three females with performance or broadcasting experience, who have not had a cold recently and speak Mandarin accurately, to record their voices. The recording is conducted in a quiet room with no echo, and the performers are in separate booths, while the recording staff are outside the booths and cannot see the performers' facial expressions and movements, only their voices. Before recording, the performers are told to speak in their own emotional expression. They can make whatever facial expressions and movements they want, as long as they do not make any noise that will interfere with the recording.

The recording was in mono, with 16-bit quantization and a sampling rate of 11025Hz. Each word and short phrase should be able to express six types of emotions, as shown in Table 1. Each long paragraph is composed of 4–5 or 5–6 short sentences. Long paragraph always contains some emotions or neutral emotions to some extent. All selected long paragraphs have certain emotions.

Listening test is conducted for verification of the emotion annotations. The work flow of data annotation is shown in Fig. 4. Listening tests are commonly used for verifying the accuracy of emotion annotations in speech datasets, and they are an important step in developing and evaluating speech emotion recognition systems. Human listeners are asked to listen to audio samples and provide their own annotations of the emotions expressed in the speech. These annotations are then compared to the existing annotations in the dataset, and any discrepancies or errors can be identified and corrected.

We constructed a number of statistic features for emotion recognition. As shown in Table 2, examples of statistical features include MFCCs, pitch, formant features. These features are derived from the acoustic properties of the speech signal and can provide information about the emotional state of the speaker.

**Fig. 4** Flowchart of the data annotation for emotional speech



**Table 2** Statistic feature set used for emotion analysis

Feature index	Feature description	Feature type
1–15	Mean, maximum, minimum, range, std of pitch, first-order difference, second-order difference	Prosodic features
16–90	Mean, maximum, minimum, range, std of first to fifth format frequency, first-order, second-order difference	Voice quality features
91–165	Mean, maximum, minimum, range, std of first to fifth format bandwidth, first-order, second-order difference	Voice quality features
166–180	Mean, maximum, minimum, range, std of short time energy, first-order, second-order difference	Prosodic features
181–375	Mean, maximum, minimum, range, std of MFCC 0 to 12 order, first-order, second-order difference	Voice quality features

### 3.2 Feature Compensation Results

Before and after the feature compensation, the feature dimension of the sample produces a clear improvement. The input feature sequence is mapped into the compensated feature sequence through the neural network, and its difference in sample gender is reduced. As shown in Tables 3 and 4, we can see happiness is improved around 3 percentage, and neutral is improved around 2 percentage.

In Table 5, the presented results serve as a basis for comparing the effectiveness of feature compensation by contrasting the metrics before and after compensation. The evaluation metrics used are precision, recall, and F1 score, which assess the model's performance in classifying different emotion classes. Before compensation, the model

**Table 3** Emotion recognition results before feature compensation (SEU database)

Ground truth	Recognition results (%)					
	Happy	Fear	Sad	Surprise	Angry	Neutral
Happy	<b>71.1</b>	5.2	6.3	7.5	3.8	6.1
Fear	5.3	<b>74.9</b>	6.4	5.5	2.0	5.9
Sad	3.1	3.2	<b>79.1</b>	4.5	6.7	3.4
Surprise	7.1	3.5	4.0	<b>75.3</b>	4.5	5.6
Angry	5.6	3.4	2.1	5.1	<b>75.1</b>	8.7
Neutral	8.4	6.5	4.4	4.5	5.7	<b>70.5</b>

The bold indicate the percentages of the correctly recognized emotion samples

**Table 4** Emotion recognition results after feature compensation (SEU database)

Ground truth	Recognition results (%)					
	Happy	Fear	Sad	Surprise	Angry	Neutral
Happy	<b>74.2</b>	4.4	5.1	7.5	3.6	5.2
Fear	4.5	<b>75.7</b>	5.1	5.9	2.1	6.7
Sad	3.4	3.5	<b>81.9</b>	4.2	3.1	3.9
Surprise	7.4	3.1	4.3	<b>77.1</b>	3.7	4.4
Angry	5.1	3.2	2.1	5.5	<b>78.4</b>	5.7
Neutral	7.2	6.1	4.5	4.2	5.4	<b>72.6</b>

The bold indicate the percentages of the correctly recognized emotion samples

demonstrates moderate performance across most emotion classes, with varying levels of precision, recall, and F1 score.

Among the emotion classes evaluated, the highest performance is observed in the “Sad” class with a precision of 0.7732, recall of 0.791, and an F1 score of 0.7820. This indicates that the model exhibits relatively good capability in identifying instances of sadness, demonstrating a balanced precision and recall. On the other hand, the lowest performance is observed in the “Neutral” class with a precision of 0.7036, recall of 0.705, and an F1 score of 0.7043. Although the model shows a reasonably balanced precision and recall for neutral instances, there is room for improvement to achieve higher accuracy.

The results indicate areas for improvement, such as enhancing recall rates to reduce missed emotions. By comparing these metrics before and after compensation, we can determine the effectiveness of the feature compensation technique in enhancing the model’s emotion classification capabilities.

As shown in Table 6, the results after feature compensation demonstrate the effectiveness of the method in enhancing the emotion classification model’s performance. The two highest performing emotion classes are “Angry” and “Sad.” The “Angry” class exhibits the highest precision (0.8141), recall (0.784), and F1 score (0.7988) after compensation, indicating a significant improvement in accurately identifying instances of anger. The second highest performing class, “Sad,” also shows notable enhancements in precision (0.7951), recall (0.819), and F1 score (0.8069), further

**Table 5** Evaluation metrics for each emotion class before compensation (SEU database)

Emotion class	Precision	Recall	F1 score
Happy	0.7068	0.711	0.7089
Fear	0.7746	0.749	0.7616
Sad	0.7732	0.791	0.7820
Surprise	0.7354	0.753	0.7441
Angry	0.7679	0.751	0.7594
Neutral	0.7036	0.705	0.7043

**Table 6** Evaluation metrics for each emotion class after compensation (SEU database)

Emotion class	Precision	Recall	F1 score
Happy	0.72888016	0.742	0.73538157
Fear	0.78854167	0.757	0.77244898
Sad	0.79514563	0.819	0.80689655
Surprise	0.73850575	0.771	0.75440313
Angry	0.81412253	0.784	0.79877738
Neutral	0.73705584	0.726	0.73148615

validating the effectiveness of the compensation method. These results highlight the success of the feature compensation technique in improving the model's ability to recognize and classify emotions, as reflected by the enhanced evaluation metrics.

Similar improvements can be observed on EMO-DB, as shown in Tables 7 and 8. Neutral is improved 2 percentage, and fear is improved around 2 percentage. We can see from Tables 9 and 10. It is evident that there has been an improvement in the feature compensation. After compensation, we can observe higher values for precision, recall, and F1 score across most emotion classes compared to the results before compensation. For instance, the precision for Happy increased from 0.9151 to 0.9241, recall increased from 0.851 to 0.864, and F1 score increased from 0.8819 to 0.8930. Similarly, improvements are seen in other emotion classes as well. This enhancement suggests that the feature compensation technique implemented in the evaluation has resulted in better accuracy and performance for emotion classification in the EMO-DB dataset.

In Figs. 5 and 6, it displays a comparison of precision, recall, and F1 metrics for emotion class recognition before and after compensation. The left subplot illustrates the performance metrics before compensation. The bars depict the precision, recall, and F1 scores for each emotion class. The legend provides a clear distinction of the metrics' color-coded representation. The right subplot showcases the corresponding metrics after compensation. Notably, after compensation, improvements are observed in precision, recall, and F1 scores across various emotion classes. The data highlights the effectiveness of the compensation approach in enhancing emotion class recognition performance in both databases. These findings contribute valuable insights to the field of emotion recognition and can aid in developing more accurate and reliable emotion recognition classifiers.

**Table 7** Emotion recognition results before feature compensation (EMO-DB database)

Ground truth	Recognition results (%)						
	Happy	Disgust	Fear	Sad	Boredom	Angry	Neutral
Happy	<b>85.1</b>	3.4	1.5	0.8	2.1	2.5	4.6
Disgust	0	<b>91.1</b>	3.3	2.1	1.2	0.3	2.0
Fear	4.5	4.1	<b>84.3</b>	2.3	1.4	0.3	3.1
Sad	1.1	2.1	3.1	<b>90.0</b>	1.1	0.2	2.4
Boredom	0.1	2.1	4.1	1.1	<b>91.0</b>	1.1	0.5
Angry	0.1	0.2	1.1	3.1	2.1	<b>91.2</b>	2.2
Neutral	2.1	2.3	4.1	1.3	4.1	4.5	<b>81.6</b>

The bold indicate the percentages of the correctly recognized emotion samples

**Table 8** Emotion recognition results after feature compensation (EMO-DB database)

Ground truth	Recognition results (%)						
	Happy	Disgust	Fear	Sad	Boredom	Angry	Neutral
Happy	<b>86.4</b>	3.0	1.5	0.4	2.2	2.2	4.3
Disgust	0	<b>92.1</b>	3.4	1.2	1.1	0.5	1.7
Fear	3.3	4.2	<b>86.4</b>	2.1	1.5	0.2	2.3
Sad	1.3	0.3	3.2	<b>92.1</b>	1.5	0.4	1.2
Boredom	0.3	2.2	2.4	0.1	<b>92.5</b>	2.0	0.5
Angry	0	0.3	1.0	0.3	2.4	<b>94.5</b>	1.5
Neutral	2.2	3.1	3.2	2.4	3.0	2.5	<b>83.6</b>

The bold indicate the percentages of the correctly recognized emotion samples

**Table 9** Evaluation metrics for each emotion class before compensation (EMO-DB)

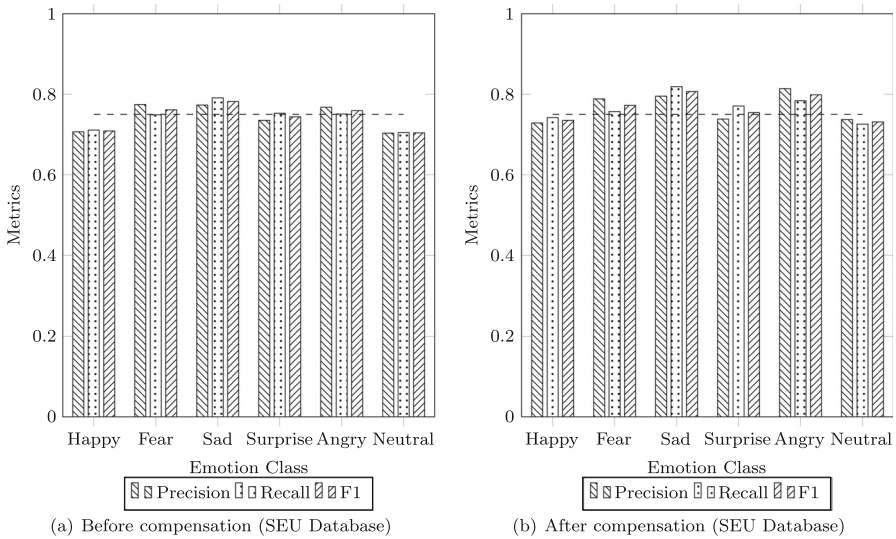
Emotion class	Precision	Recall	F1 score
Happy	0.9151	0.851	0.8819
Disgust	0.8651	0.911	0.8875
Fear	0.8305	0.843	0.8367
Sad	0.8937	0.9	0.8969
Boredom	0.8835	0.91	0.8966
Angry	0.9111	0.912	0.9115
Neutral	0.8465	0.816	0.8310

### Comparison Between Denoising Autoencoder and Existing Feature Compensation Algorithm

Feature compensation or normalization methods are employed to standardize the features or variables within a dataset, ensuring they are on a consistent scale or distribution. These techniques are valuable for enhancing the performance of machine learning algorithms and ensuring that all features contribute proportionately to the analysis.

**Table 10** Evaluation metrics for each emotion class after compensation (EMO-DB)

Emotion class	Precision	Recall	F1 score
Happy	0.9241	0.864	0.8930
Disgust	0.8755	0.921	0.8977
Fear	0.8546	0.864	0.8593
Sad	0.9341	0.921	0.9275
Boredom	0.8877	0.925	0.9060
Angry	0.9238	0.945	0.9343
Neutral	0.8791	0.836	0.8570

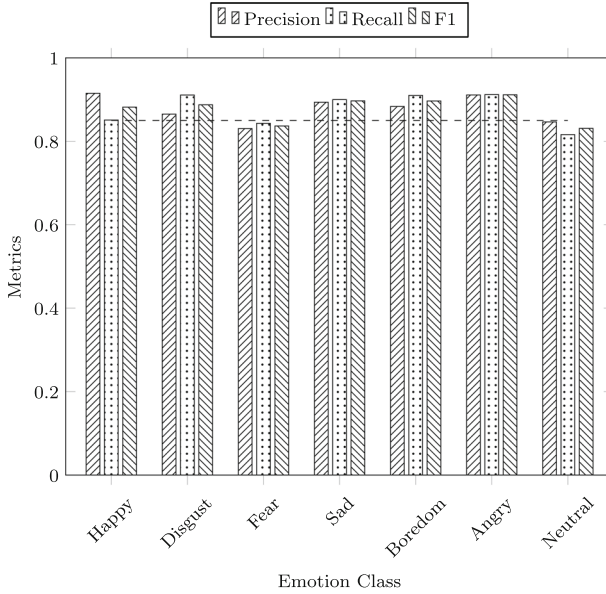


**Fig. 5** Comparison of precision, recall, and F1 for emotion class recognition before and after compensation (SEU Database)

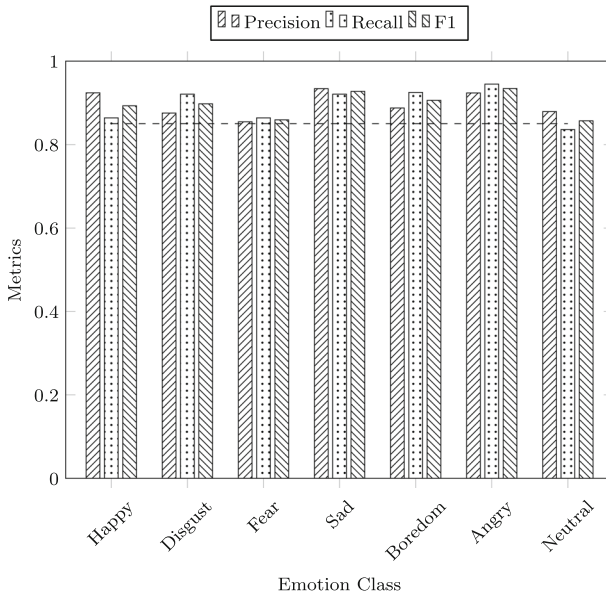
Mean shift [27] can be utilized for feature compensation by applying it to the feature space clustering. This approach aims to shift the feature distribution towards a desired target distribution, thereby facilitating feature normalization or compensation. Subsequently, normalization and standardization can be performed within each arbitrary cluster shape to enhance the data distribution for modeling purposes.

As shown in Figs. 7 and 8, the denoising autoencoder exhibits higher F1 scores of 0.735 (Happy), 0.772 (Fear), 0.807 (Sad), 0.754 (Surprise), 0.799 (Angry), and 0.731 (Neutral). In comparison, the mean-shift method achieved slightly lower F1 scores: 0.713 (Happy), 0.764 (Fear), 0.789 (Sad), 0.751 (Surprise), 0.763 (Angry), and 0.705 (Neutral). These results indicate that the denoising autoencoder approach outperforms the mean-shift method in improving feature compensation based on the SEU dataset.

Similarly, on EMO-DB, the data represents the F1 scores for emotion classes such as Happy, Disgust, Fear, Sad, Boredom, Angry, and Neutral, using the same denois-



(a) Before compensation (EMO-DB Database)



(b) After compensation (EMO-DB Database)

**Fig. 6** Comparison of precision, recall, and F1 for emotion class recognition before and after compensation (EMO-DB Database)

ing autoencoder and mean-shift methods. The denoising autoencoder method yields higher F1 scores. In contrast, the mean-shift method achieves slightly lower F1 scores.



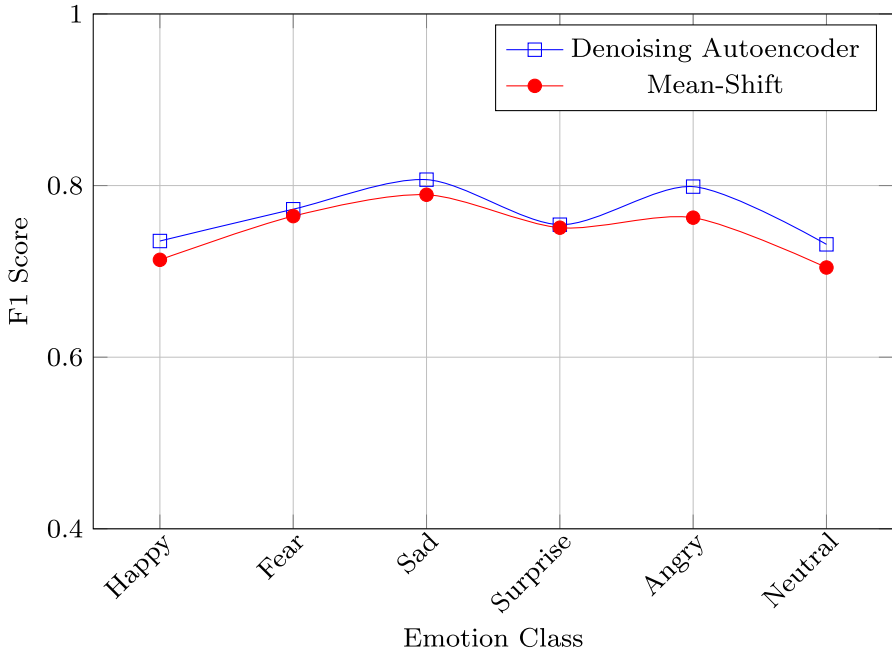


Fig. 7 Denoising autoencoder-based feature compensation (SEU)

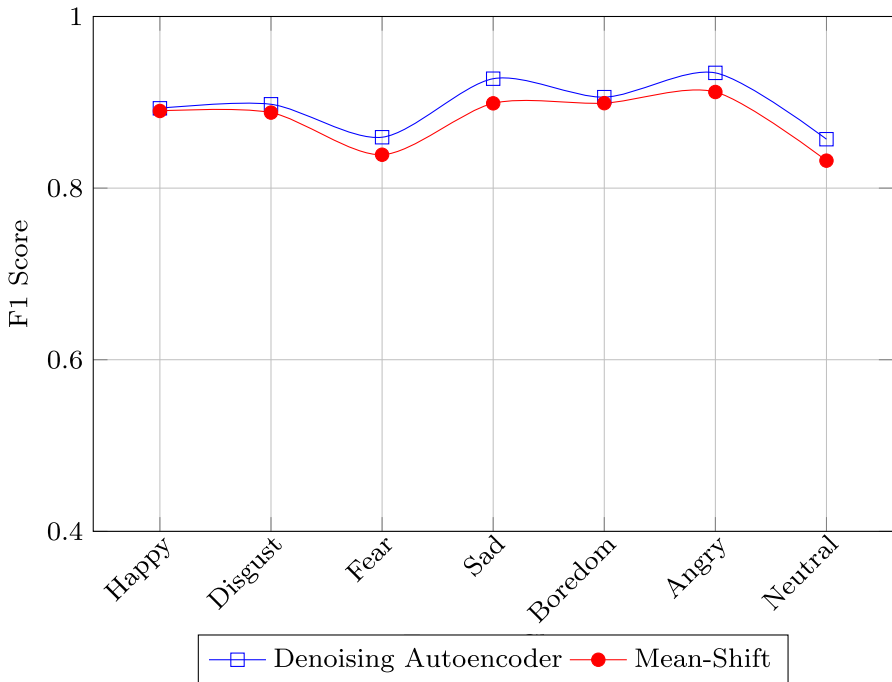
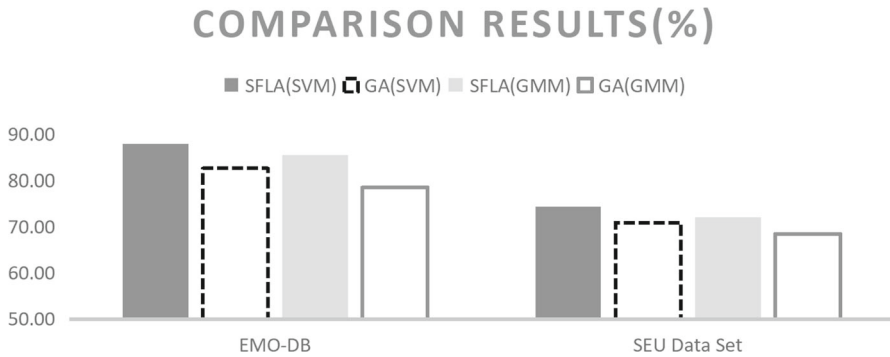


Fig. 8 Denoising autoencoder-based feature compensation (EMO-DB)



**Fig. 9** Comparison of SFLA and GA optimization results

These results highlight that the denoising autoencoder approach demonstrates superior performance over the mean-shift method in enhancing feature compensation on both the SEU dataset and EMO-DB dataset.

Overall, we can observe that the denoising autoencoder method exhibits better improvement in feature compensation compared to the mean-shift method, as evidenced by the higher F1 scores achieved across various emotion classes in both datasets.

### 3.3 Feature Optimization Results

We compare the GA and SFLA algorithms and optimize the combination of features to improve the emotion recognition results at the frame level.

In Fig. 9, we compared the average recognition rates for EMO-DB and SEU Data Set. SFLA results are better than GA, with better frame-level recognition rates.

In our experiment, we use GMM and SVM to construct the classifiers. As we focus on feature optimization, we use GMM and SVM as two typical classifiers for emotion recognition to verify the feature optimization method. It can be seen that SVM gives better average recognition rates. In the subsequent experiments, we use SVM for long-time speech emotion recognition experiments.

In the subsequent experiments, we use the feature combination given by SFLA, as shown in Tables 11 and 12. The feature optimization results using GA are provided in Tables 13 and 14.

The population size for GA is set to 200 as the starting point. The rank-based selection method is used. The crossover rate is set to 0.5 and the mutation rate is set to 0.001 as the starting point. The population size for SFLA is set to 100 as the starting point. The mutation rate is set to 0.01. The crossover rate is set to 0.5. Detailed information is summarized in Tables 15 and 16.

**Table 11** Optimized combination of speech emotional features using SFLA(SEU dataset)

Index	Feature description	Feature category
1	Std of pitch frequency	Prosodic feature
2	Min of first-order difference of short-time energy	Prosodic feature
3	Std of MFCC2	Voice quality feature
4	Mean of F1 frequency	Voice quality feature
5	Mean of F3 bandwidth	Voice quality feature
6	Std of MFCC1	Voice quality feature
7	Std of short-time energy	Prosodic feature
8	Min of short-time energy	Prosodic feature
9	Std of first-order difference of short-time energy	Prosodic feature
10	Mean of first-order F1 frequency	Voice quality feature

**Table 12** Optimized combination of speech emotional features using SFLA(EMO-DB)

Index	Feature description	Feature category
1	Mean of first-order difference of short-time energy	Prosodic feature
2	Std of first-order difference of short-time energy	Prosodic feature
3	Min of short-time energy	Prosodic feature
4	Mean of F1 frequency	Voice quality feature
5	Mean of first-order difference F3 frequency	Voice quality feature
6	Std of MFCC1	Voice quality feature
7	Std of pitch frequency	Prosodic feature
8	Std of short-time energy	Prosodic feature
9	Mean of MFCC2	Voice quality feature
10	Std of MFCC4	Voice quality feature

**Table 13** Optimized combination of speech emotional features using GA (SEU dataset)

Index	Feature description	Feature category
1	Mean of bandwidth of F3	Voice quality feature
2	Std of MFCC1	Voice quality feature
3	Min of short-time energy	Prosodic feature
4	Mean of first-order difference of short-time energy	Prosodic feature
5	Mean of MFCC2	Voice quality feature
6	Std of MFCC4	Voice quality feature
7	Std of pitch frequency	Prosodic feature
8	Std of short-time energy	Prosodic feature
9	Mean of short-time energy	Prosodic feature
10	Mean of first-order F3 frequency	Voice quality feature

**Table 14** Optimized combination of speech emotional features using GA(EMO-DB)

Index	Feature description	Feature category
1	Std of pitch frequency	Prosodic feature
2	Mean of first-order difference of short-time energy	Prosodic feature
3	Std of MFCC5	Voice quality feature
4	Mean of MFCC5	Voice quality feature
5	Mean of first-order F1 frequency	Voice quality feature
6	Std of short-time energy	Prosodic feature
7	Min of short-time energy	Prosodic feature
8	Std of first-order F3 frequency	Voice quality feature
9	Mean of first-order F3 frequency	Voice quality feature
10	Mean of bandwidth of F1	Voice quality feature

**Table 15** Parameter settings for genetic algorithm (GA) in feature selection

Parameter	Description	Value
Population size	Number of individuals in each generation	200
Number of generations	Total number of generations	200
Crossover rate	Probability of crossover occurring	0.5
Mutation probability	Probability of mutation occurring	0.001

**Table 16** Parameter settings for shuffled frog-leaping algorithm (SFLA) in feature selection

Parameter	Description	Value
Population size	Number of frogs in the population	100
Number of generations	Total number of generations	100
Number of memeplexes	Number of subpopulations (memeplexes)	5
Number of frogs per memeplex	Number of frogs in each memeplex	5
Shuffling factor	Fraction of memeplex frogs to be shuffled	0.6
Crossover probability	Probability of crossover occurring	0.5
Mutation probability	Probability of mutation occurring	0.01
Local search iterations	Number of iterations for local search optimization	100

### 3.4 Long-Time Emotion Recognition Results

Using our proposed method, considering the accentuation, the context and the long-time dependency, speech emotion recognition results can be further improved.

The transformer model is configured with various parameter settings to optimize its performance during training and inference. A dropout rate of 0.1 is applied to the model, randomly deactivating 10% of the neurons during training to prevent overfit-

**Table 17** Emotion recognition results using long-time emotion fusion (SEU database)

Ground truth	Recognition results (%)					
	Happy	Fear	Sad	Surprise	Angry	Neutral
Happy	<b>79.4</b>	4.1	3.2	5.1	3.8	4.4
Fear	3.3	<b>77.2</b>	5.1	3.4	6.1	4.9
Sad	2.5	3.1	<b>84.5</b>	4.4	2.1	3.4
Surprise	5.1	3.3	4.2	<b>80.5</b>	3.1	3.8
Angry	5.2	3.2	1.0	5.7	<b>81.7</b>	3.2
Neutral	6.1	5.1	4.4	3.8	2.8	<b>77.8</b>

The bold indicate the percentages of the correctly recognized emotion samples

**Table 18** The final evaluation metrics for each emotion class

Emotion class	Precision	Recall	F1 score
Happy	0.7815	0.794	0.7877
Fear	0.8042	0.772	0.7878
Sad	0.8252	0.845	0.8350
Surprise	0.7823	0.805	0.7935
Angry	0.8203	0.817	0.8186
Neutral	0.7979	0.778	0.7878

ting. The learning rate is set to 0.001, controlling the step size of parameter updates during optimization. A batch size of 64 is utilized, determining the number of training examples processed together in each iteration. The maximum sequence length is limited to 512, ensuring efficient processing and memory utilization. Finally, the model undergoes 10 training epochs, indicating the number of times the entire training dataset is processed. These parameter settings collectively define the behavior and capacity of the transformer model, allowing it to effectively process and learn from sequential data, such as long-time speech emotion sequences.

As shown in Tables 17 and 18, the final results tested on SEU database show that the proposed methods resulted a promising performance for long-time speech emotion recognition. The recognition accuracy is further improved. In Figs. 10 and 11, we further illustrated the improvement using our proposed algorithms.

As shown in Fig. 12, we compared the conventional averaged weights fusion with the proposed accentuation-based approach. In the context of long-time emotional speech analysis, two fusion techniques that can be employed are averaged weights fusion and accentuation-based fusion. These techniques aim to combine multiple sources of information or features to enhance the overall performance of emotion recognition systems.

Averaged weights fusion involves assigning equal importance to all the input features or sources, which is conventionally used when not considering the different characters in long-time emotional speech. Each feature is weighted equally, and their contributions are averaged to obtain a combined representation. This fusion method assumes that all features are equally relevant and can provide valuable information for

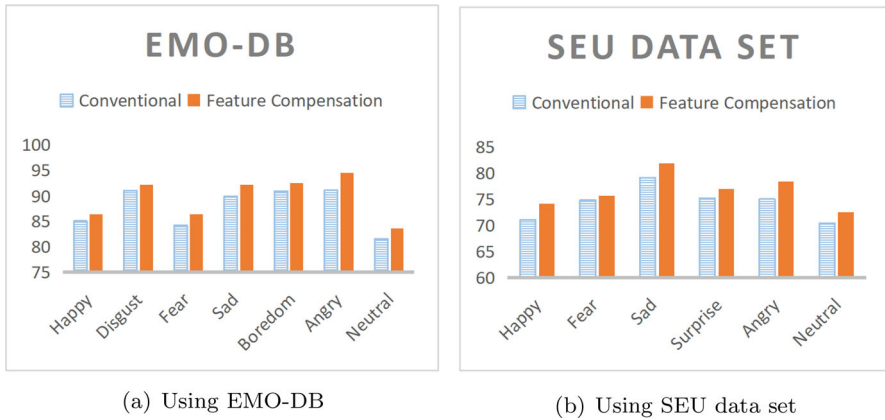


Fig. 10 Improvement of the emotion recognition results using the feature compensation algorithm

### LONG-TIME SPEECH EMOTION

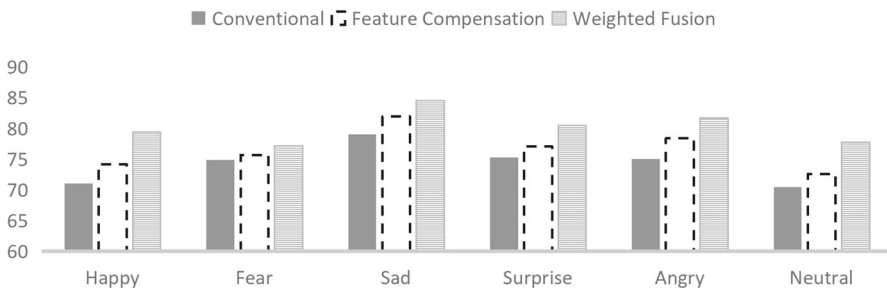


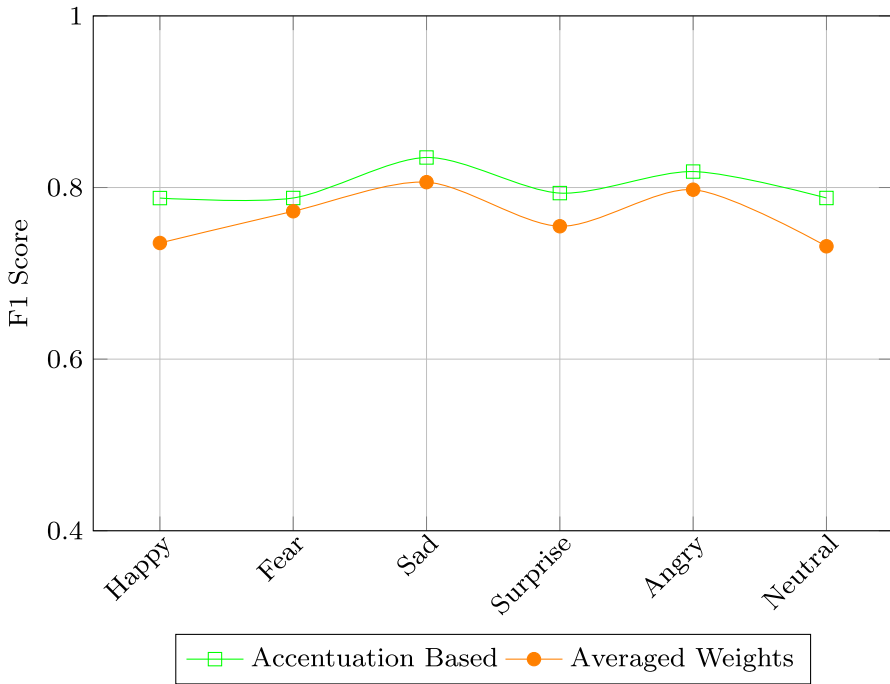
Fig. 11 Improvement of the emotion recognition results using the long-time emotion fusion algorithm

emotion recognition. However, it may not consider the varying importance of different features in capturing emotional cues in long-time speech.

From Fig. 12, we can see that the green line represents the accentuation-based method, and the orange line represents the averaged weights method. Both methods have F1 scores for each emotion class plotted for comparison.

We can observe that the F1 scores vary for different emotion classes and between the two methods. In general, the “Sad” emotion class has the highest F1 score for both methods, followed by “Angry” and “Surprise”. The lowest F1 scores are seen for the “Neutral” emotion class.

The proposed accentuation-based method constantly has higher F1 scores compared to the averaged weights method across the emotion classes. We further calculated the p-value of significance equal to 0.0061, which is smaller than the common alpha level (significance level), 0.05 or 0.01. Based on the given information, the proposed method is better than the conventional method in long-time speech emotion recognition, as it achieves marginally higher F1 scores.



**Fig. 12** Comparison with the existing method in long-time speech emotion recognition

## 4 Discussion

Speech emotion recognition is a challenging task because it requires not only understanding the words being spoken but also the emotional content conveyed by the speaker. One of the key factors that can affect the accuracy of speech emotion recognition is the context in which the speech is spoken. Understanding the context of the speech can help improve the accuracy of emotion recognition by providing additional information about the speaker's intentions and emotional state.

We can improve the accuracy of speech emotion recognition using the long-time dependency of emotions. Emotions are often expressed over an extended period, and they can change rapidly, making it difficult to accurately capture and classify them. To address this, we have explored various techniques such as feature compensation, feature optimization, and fusion of recognition results in long-time speech.

## 5 Conclusion

In this paper, we studied long-time speech emotion recognition, which is an important topic in real-world applications, yet lack of systematic and in-depth research. First, we used denoising autoencoder for emotion feature compensation and then used SFLA for feature selection. Second, we studied accentuation-based emotion fusion, and we used

transformer to predict the probability of the next emotion, and corrected errors from the view of emotion sequence. Third, we verified our methods on two different databases. The feature compensation and optimization are tested and compared on both databases, and the long-time emotion fusion and recognition is tested on a local database. The results show that our framework is suitable for emotional feature extraction and long-time emotion recognition.

In our future work, we will consider more factors related to long-time emotional behavior and study different context information in paragraph level. Future work will emphasize the analysis of long-term emotional behavior and the exploration of context at the paragraph level. This approach provides a more comprehensive understanding of emotions and their dynamics, going beyond traditional sentence-level analysis. By considering factors that influence emotional behavior over time and incorporating paragraph-level context, researchers can uncover patterns, trends, and changes in emotions within specific contexts. This research has practical applications in areas such as mental health, customer experience, and education.

**Data Availability** The EMO-DB database is the freely available German emotional database. EMO-DB can be downloaded publicly from <https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emo-db>.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest related to this work.

## References

1. A.A. Abdelhamid, E.S.M. El-Kenawy, B. Alotaibi, G.M. Amer, M.Y. Abdelkader, A. Ibrahim, M.M. Eid, Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. *IEEE Access* **10**, 49265–49284 (2022)
2. S. Akinpelu, S. Viriri, Robust feature selection-based speech emotion classification using deep transfer learning. *Appl. Sci.* **12**(16), 8265 (2022)
3. F. Albu, D. Hagiescu, L. Vladutu, M.A. Puica, Neural network approaches for children’s emotion recognition in intelligent learning applications, in: *International Conference on Education and New Learning Technologies*, 3229–3239 (2015)
4. S.B. Alex, L. Mary, B.P. Babu, Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. *Circuits Syst. Signal Process.* **39**(11), 5681–709 (2020)
5. T. Anvarjon, S. Kwon, Deep-net: a lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors* **20**(18), 1–16 (2020)
6. B.T. Atmaja, A. Sasou, M. Akagi, Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **140**, 11–28 (2022)
7. G. Choudhary, R. Meena, K. Mohbey, Speech emotion based emotion recognition using deep neural networks. *J. Phys. Conf. Ser.* **2236**(1), 012003 (2022)
8. A. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci.* **114**(38), E7900–E7909 (2017)
9. M.S. Fahad, A. Deepak, G. Pradhan, J. Yadav, DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. *Circuits Syst. Signal Process.* **40**, 466–89 (2021)
10. C. Fu, Q. Deng, J. Shen, H. Mahzoon, H. Ishiguro, A preliminary study on realizing human–robot mental comforting dialogue via sharing experience emotionally. *Sensors* **22**(3), 991 (2022)
11. I. Gat, et al., Speaker normalization for self-supervised speech emotion recognition, in: *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7342–7346 (2022)



12. N. Hajarolasvadi, H. Demirel, 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. *Entropy* **21**(5), 479 (2019)
13. C. Huang, B. Song, L. Zhao, Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering. *Int. J. Speech Technol.* **19**(4), 805–816 (2016)
14. C. Huang, Y. Jin, Q. Wang, Speech emotion recognition based on decomposition of feature space and information fusion. *J. Signal Process.* **26**(6), 835–842 (2010)
15. C. Huang, Y. Jin, Y. Zhao, Y. Yu, L. Zhao, Speech emotion recognition based on re-composition of two-class classifiers, in: *The 3rd International conference on affective computing and intelligent interaction and workshops* (2009)
16. Y. Jin, C. Huang, L. Zhao, A semi-supervised learning algorithm based on modified self-training SVM. *J. Comput.* **6**(7), 1438–1443 (2011)
17. S.R. Kadiri, P. Gangamohan, S.V. Gangashetty, P. Alku, B. Yegnanarayana, Excitation features of speech for emotion recognition using neutral speech as reference. *Circuits Syst. Signal Process.* **39**(9), 4459–81 (2020)
18. B. Maji, M. Swain, Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features. *Electronics* **11**(9), 1328 (2022)
19. K. Manohar, E. Logashanmugam, Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. *Knowl. Based Syst.* **246**, 108659 (2022)
20. M. Oaten, R.J. Stevenson, T. Case, Disgust as a disease-avoidance mechanism. *Psychol. Bull.* **135**(2), 303–321 (2009)
21. T. Özseven, A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **146**, 320–326 (2019)
22. L. Pandey, R.M. Hegde, Keyword spotting in continuous speech using spectral and prosodic information fusion. *Circuits Syst. Signal Process.* **38**, 2767–91 (2019)
23. V.M. Praseetha, P.P. Joby, Speech emotion recognition using data augmentation. *Int. J. Speech Technol.* **25**(4), 783–792 (2022)
24. H. Saad, F. Mahmud, M. Shaheen, M. Hasan, P. Farastu, M. Kabir, Is speech emotion recognition language-independent? Analysis of English and Bangla languages using language-independent vocal features. *arXiv preprint, arXiv:2111.10776* (2021)
25. C. Wu, C. Huang, H. Chen, Text-independent speech emotion recognition using frequency adaptive features. *Multimed. Tools Appl.* **77**(18), 24353–24363 (2018)
26. X. Xu et al., Graph learning based speaker independent speech emotion recognition. *Adv. Electr. Comput. Eng.* **14**(2), 17–23 (2014)
27. L. You, H. Jiang, J. Hu, C. H. Chang, L. Chen, X. Cui, M. Zhao, GPU-accelerated faster mean shift with Euclidean distance metrics, in: *2022 IEEE 46th Annual Computers, Software, and Applications Conference*, 211–216 (2022)
28. X. Zhang et al., Recognition of practical speech emotion using improved shuffled frog leaping algorithm. *Chin. J. Acoust.* **33**(4), 441–441 (2014)
29. C. Zou, C. Huang, D. Han, L. Zhao, Detecting practical speech emotion in a cognitive task, in: *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, 1–5 (2011)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.