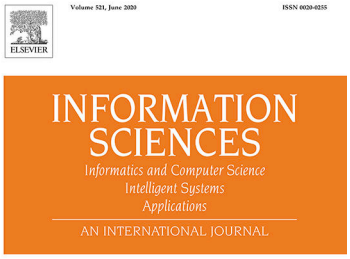
Journal Pre-proofs



Exploiting the Potentialities of Features for Speech Emotion Recognition

Dongdong Li, Yijun Zhou, Zhe Wang, Daqi Gao

PII: S0020-0255(20)30951-8
DOI: https://doi.org/10.1016/j.ins.2020.09.047
Reference: INS 15886

To appear in: *Information Sciences*

Available online at www.sciencedirect.com
ScienceDirect

Received Date: 29 September 2019
Accepted Date: 19 September 2020

Please cite this article as: D. Li, Y. Zhou, Z. Wang, D. Gao, Exploiting the Potentialities of Features for Speech Emotion Recognition, *Information Sciences* (2020), doi: https://doi.org/10.1016/j.ins.2020.09.047

Exploiting the Potentialities of Features for Speech Emotion Recognition

Dongdong Li, Yijun Zhou, Zhe Wang*, Daqi Gao

Abstract

In recent years, studies on speech signals have increasingly paid attention to emotional information. The most challenging aspect in speech emotion recognition (SER) is choosing the optimal speech feature representation. According to the statistical analysis, the roles of each speech feature differ under different emotions, indicating that different features have different abilities in distinguishing emotions. This study proposes an emotional-category based feature weighting (ECFW) method, which aims at finding the prominence of each feature under different emotions and applying this prominence as priori knowledge. Furthermore, previous studies have paid little attention to matching the relationship between speech features and models. This study argues that different combinations of models and features result in large differences in the performance of SER, which are evaluated by several experiments. Features must be modeled with appropriate approaches to extract the most valuable information for emotional representation. Then, the best combinations of features and models are selected to test our method. The method is applied on three commonly used speech emotion databases, IEMOCAP, MASC, and EMO-DB. The results show that ECFW significantly improves the performance of SER tasks.

Index Terms

Feature Optimization, Feature Selection, Speech Emotion Recognition, Deep Learning

I. INTRODUCTION

Speech signal contains abundant information that extends the content of a written message by including factors such as the identity of the speaker, their emotional state, information status, and intonational patterns [22]. Typically, the speech recognition system explores speech features

*Corresponding author. E-mail: wangzhe@ecust.edu.cn (Z. Wang).

Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai, 200237, P.R. China

behind the speech signal and identifies their content using algorithms. Since the late 1950s, numerous studies were conducted on speech recognition, and they achieved significant progress [31]. At the same time, to obtain a better human–computer interaction experience, empowering machines with emotional expressive ability and making them capable of recognizing the human emotional state has gained increasing popularity in the fields of human–computer interaction. This led to a new research field of speech emotion recognition (SER). SER is considered as one of the most important research areas in the past decade. Numerous researchers are attracted by the automated analysis of human affective behavior. However, despite significant efforts made by speech recognition, SER requires considerable work to achieve more natural interactions between human and machine.

In recent years, deep learning has been widely applied in various fields, naturally including SER [24]. Deep neural networks (DNNs) and convolutional neural networks (CNNs) are two typical feed-forward neural networks that take the place of the traditional structures and are often used as the main frameworks for SER [2]. Recurrent architectures, such as long short-term memory (LSTM), carry temporal information, which can also be useful in SER [40]. In contrast to the traditional methods, deep learning is considered as an end-to-end learning method, which means that it has the ability to spontaneously learn the features behind the data. In the field of SER, even unprocessed original speech signals are used to classify emotions by deep learning architecture. In [47], the author compared the emotional classification between the original speech signal and log-mel spectrograms using the 1D & 2D CNN LSTM networks, respectively. The experimental results show that the emotional classification ability of the original speech signal is inferior to the extracted speech features, indicating that features have a significant influence on SER.

In fact, the factors that affect a person's emotion are complex and varied. Individuals experience various psychological changes under different emotional states. These changes cause them to attach emotional fluctuations to their speech, thus providing emotional information, which is key to speech emotion recognition. The speech features are extracted to describe this emotional information. Many speech features serve to distinguish between different emotions. For example, the pitch of a fearful or angry voice is significantly higher than that of a neutral or disgusted voice [35]. High arousal emotions such as anger, happiness, or surprise all yield increased energy, while disgust and sadness result in decreased energy [25]. The F0 contour decreases over time during the emotion of anger, while it increases over time during the emotion of happiness

[17]. Further, anger lasts shorter than sadness [8]. For a sad voice, the standard deviation of loudness is higher, while the opposite trend is noted for a happy voice [48]. Therefore, the distribution of specific speech features under different emotions is diverse, and this potential information is a useful supplement to the SER task. Thus, to fully utilize this information, we propose an emotional-category based feature weighting (ECFW) method to exploit this potential by calculating the prominence of different speech features under different emotions. First, we calculate the prominence of each feature under diverse emotions by correlation algorithms. Then, we assign the prominence value to each feature to enhance their differences. Finally, the special learning method based on deep learning is employed to classify the newly generated feature sets. However, one problem persists in this approach, namely, determining which speech features must be used as initial input.

To solve this problem, the matching relationship between features and models is studied. According to numerous studies [12], a variety of speech features exists. Some of them are based on frame-level extraction, such as pitch, energy, spectral features, and TEO-based features. Others are based on statistical functions, such as the mean, maximum, variance, and derivative of various speech features. The former are often referred to as handcrafted low-level descriptors (LLDs), and the latter are high-level statistic features (HSFs). Hence, in this study, we experiment with both LLDs and HSFs on different deep learning models to find a set of well-adapted feature combinations. Further, considering the high dimensions of speech features, we compare several feature selection algorithms and obtain a more appropriate feature representation [29].

In the studies on SER, there is no generally accepted set of features for precise and distinctive classification. Speech features are usually selected merely based on the researchers' experience, and processed with equal weights during modeling. However, studies of speech features show that different speech features have a variety of ranges and trends under different emotional states, i.e., each feature has its unique identity under different emotions. For speech emotion recognition, it is crucial to make better use of the variety of identities in presenting different emotions. Thus, a new algorithm, ECFW, is introduced in this article. Apart from previous modeling processes, where all features are assigned fixed weights, ECFW assigns different weights to the features based on emotion classifications. The identities of different emotions are retrieved through calculation of the relevance of features and strengthened by assigning corresponding weights to improve the accuracy of emotion recognition. Further, to obtain the optimal combination of features and models as the benchmark system, this study investigates the performance of SER under the

combination of different speech features and deep models as well as different feature selection algorithms. The main contributions of this study are summarized as follows.

1) An emotional-category based feature weighting (ECFW) method is proposed. In contrast to the traditional SER methods, which process speech features at equal weights, ECFW exploits speech features by calculating the prominence of different speech features under diverse emotions, and fully utilizing this prominence to enhance the differences between features and emotions.
2) The relationship between speech features and deep models is critically explored. This study argues that different combinations of models and features result in large differences in the performance of speech emotion recognition(SER), which are evaluated by several experiments. Optimal features that correspond to each deep model are identified, and they can be used as a reference for the feature and model selection in SER.
3) The effects of the feature selection algorithm combined with speech features are investigated based on the deep model. The most discriminative feature subsets is selected to form a more approriate feature representation as well as to reduce the feature dimensions.

The remainder of this paper is structured as follows. Section II presents related studies. Section III introduces the framework of the proposed method. The study on feature engineering for SER is presented in Section IV. Section V discusses the details of the proposed method. Our experiments and results are explained in Section VI, and Section VII provides the conclusion.

II. Related Work

In the past few years, numerous studies conducted on SER have focused on the improvement and integration of deep models. In [39], a method based on the combination of a two-layer CNN and several extended speech features was used to recognize the emotion from the EmotAsS data set. Luo et al. proposed an HSF-CRNN SER system, which integrated the handcrafted HSFs and CRNN-learned feature to obtain a joint representation of speech on the IEMOCAP data set[26]. An Attention-BLSTM-RNNs was proposed for modeling spatio-temporal dynamics for SER by combining CNN and BLSTM[49]. Research was conducted on the integration of deep models, and some methods aimed at improving them. Neumann et al. applied the attention mechanism to CNN to explore the impact of speech features and signal length on the SER[28]. To recognize the emotion in music, Schmidt et al. presented a three-layer regression-based deep belief network, which directly learned from spectral features[13]. Mirsamadi et al. introduced a local attention

mechanism into a BLSTM network. The network focused on the most significant specific area of emotion to distinguish the most important speech representation in raw signal[30].

Moreover, learning speech features is essential for recognizing the speech emotion. Alonso et al. provided an algorithm learned by two prosodic features and four paralinguistic features related to the pitch and spectral energy to achieve a realtime SER system with low processing time[4]. Vegesna et al. analyzed the affects between prosodic features and emotions. The authors implemented several prosodic features and utilized linear discriminant analysis (LDA) to select the best features to modify the speech and reduce the emotional interference[41]. Wang et al. extracted MFCC, the Fourier parameter, fundamental frequency, energy, and zero-crossing rate from the speech signal to recognize seven emotions in three data sets[42]. Sun et al. proposed an SER method, namely, weighted spectral features based on Hu moments (HuWSF). The method combined Interspeech'10 feature sets comprising 1582 features and spectral features to achieve SER tasks on three data sets[38].

The above methods are used to solve the corresponding problems in the SER field from the perspective of modeling and feature learning. When selecting features, most studies use prosodic and spectral features to yield a better result, which can be further improved by adding or selecting the features. When constructing these models, convolutional or recurrent neural networks are commonly used. Similarly, this can be improved by integrating multiple models (e.g., CRNN) or introducing new mechanisms (e.g., attention mechanism). Although the methods mentioned above are useful for recognizing speech emotions, studies have rarely focused on matching the relationship between features and models, or the different capabilities of the features in recognizing different emotions. Inspired by previous reports[48], [44], this study finds that even the same feature has a different range and trend when extracted from different emotions, which means that there are invariable possibilities of finding some features with strong recognition abilities over a specific emotion. The ECFW is proposed to explore the potential expression ability of each feature in different emotional states by calculating the correlation of features, to improve the accuracy of emotion recognition. Further, to obtain the optimal initial feature set, this study investigates the matching relationship between features and models, as well as feature selection algorithms.

III. The framework of the SER

The overall framework of the SER method presented in this study is shown in **Fig. 1**. The

Fig. 1. SER method pipeline. (1) Output the optimal combination of feature and model by matching each feature and model. (2) Use the optimal feature selection algorithm to reduce the feature dimension. (3) Obtain a better feature representation to recognize emotions by calculating the weights.

SER method usually consists of three parts, namely, speech feature extraction, speech feature selection, and emotion classification. The proposed method is designed to improve the recognition of speech emotion from these three aspects. Concretely, for a speech sample X, several kinds of emotional features F such as the spectrogram and MFCC are extracted from speech signals. These features are matched with various deep models M, such as DNN and CNN. By applying different neural network models (M_1, M_2,..., M_j,..., M_T) to learn the extracted emotional features (F_1, F_2,..., F_i,..., F_S), a set of highly matched feature F_i and model M_j that can form the best combination suitable for emotion classification is selected for subsequent processing. To obtain a better feature representation, different feature selection methods such as ReliefF and MRMR are further employed to select the optimal speech feature subsets $X' = \{X'_1, X'_2, ...X'_i, ...X'_C\}$ in the d-dimension, where d represents the dimension of features, and C represents the number of emotional categories. Subsequently, the ECFW method is proposed to calculate the potentiality of the selected feature subsets X'_i for the ith emotion category. The corresponding weights $W_i = \{W_{i,1}, W_{i,2}, ..., W_{i,d}\}$ of each feature under different emotions can be obtained. The weights are combined with features to obtain a more distinct feature representation $W_i X'_i$, which serves as new data input. Finally, $W_i X'_i$ is fed into the selected model M_j to predict the emotional states. The details of each step are explained in the following sections.

IV. Feature Engineering for SER

Feature engineering is the process of creating features that enable machine learning algorithms to achieve the best performance by utilizing prior knowledge in the data domain. Feature extraction and selection are two important tasks in feature engineering. In present studies, speech features used in SER are often based on common sense and experience, and lack of research on the matching relationship between models and features. Therefore, choosing a suitable set of features to form the best match with the follow-up deep model is the key to improving the SER system.

A. *Speech Feature Extraction*

As mentioned above, speech features used to recognize emotional categories are often divided into two types: one is the feature representation referred to as LLDs, extracted by frames, and the other is the statistical computing features of LLDs called HSFs.

For LLDs, spectrum-based speech features are particularly widely used in SER [20]. Spectrum features are considered to be correlations between the changes in the vocal tract and the articulator movements. MFCC (Mel-frequency cepstral coefficients) is among the most commonly used spectral features [27]. In this study, the first 13 MFCC order coefficients and their corresponding first and second derivatives are also assumed, as in [50].

Filter bank is another commonly used spectrum-based speech feature in SER [46]. The extraction method of the filter bank is similar to that of MFCC. Meanwhile, the step of discrete cosine transform (DCT) is not required compared with MFCC. In this study, 40 filter banks are selected to form 40-dimensional filter banks features.

Apart from the two features mentioned above, some commonly used LLDs features are also adopted in this study, i.e., linear predictor cepstral coefficients (LPCC) [6], speech spectrogram [1] extracted by short-time Fourier-transform on the frame level and original speech signals as well.

The HSFs are calculated by applying a statistical function on LLDs. Compared with LLDs, HSFs have advantages of low dimensionality and fast training speed, but with little information of speech temporal [7]. There are many existing HSFs datasets, such as the Geneva minimalistic acoustic parameter set (GeMAPS) [14] and its extended version called eGeMAPS, the Interspeech'09[36], and the Interspeech'10 [37]. The GeMAPS feature set consists of 62 features, which are obtained from 18 LLDs features by statistical calculation. Further, 88 features in the eGeMAPS feature set are obtained from 25 LLDs features by statistical calculation. **Table I** lists the components of speech features in Interspeech'09 and Interspeech'10. Interspeech'09 contains 16 basic LLDs and their corresponding first-order derivatives, and it uses 12 feature statistical functions to obtain a total of 384 speech features. Interspeech'10 contains 38 basic LLDs and 38 corresponding first-order derivatives, and uses 21 statistical functions to obtain a total of 1428 features. Subsequently, 19 statistical functions (minimum value and range are removed from 21 feature statistical functions) are applied on 8 LLDs (4 LLDs of baseline frequency and their derivatives), and 152 features are obtained. Two additional features are added as the pitch

interval and the total duration, which finally amount to a total of 1582 features. In this study, these four HSFs feature sets mentioned above are all extracted and compared together to find the most suitable one for the method.

B. Speech Feature Selection

As a preprocessing step in machine learning, feature selection has achieved good performance by reducing dimensionality, removing irrelevant data, and improving comprehensibility [5]. Because the high dimensionality of speech features affects the classification performance and efficiency, the feature selection of speech features is particularly important. Feature selection methods fall into two major categories, namely, the wrapper and filter methods [29]. The difference between these two methods is that the wrapper method must consider the ability of the learning algorithm, while the filter method does not. The wrapper method is found to cause the curse of dimensionality when the dimensionality of the feature vector is too large [3]. Therefore, for high-dimensional speech emotional features, the filter method is usually chosen owing to its computational efficiency. In this study, the filter method is based on two different selection algorithms, namely, the subset search algorithm and feature-weighting algorithm.

The subset search algorithm is guided by specific evaluation indicators, and it searches the candidate feature subset to capture the optimal degree of each subset. The steps of the common

TABLE I

COMPONENTS OF SPEECH FEATURES IN INTERSPEECH'09 AND INTERSPEECH'10

	LLDs	**Statistical functions**
Interspeech'09		
	MFCC[1-12]	maxPos, minPos, amean
	Harmonics-to-Noise Ratio	max, min, mean, stddev
	Zero Crossing Rate, Pitch Frequency	lin.regc.coeff 1/2, lin.regc.coeff Q
	Root Mean Square (RMS) Frame Energy	skewness, kurtosis
Interspeech'10		
	MFCC[0-14], LSP frequency[0-7]	maxPos, minPos, amean
	Log Mel freq. band[0-7]	stddev, skewness, kurtosis
	PCM loudness	lin.regc.coeff 1/2, lin.regc.coeff A/Q
	F0 envelop, Voicing probability	quartile range (1-2)/(2-3)/(1-3), quartile 1/2/3
	JitterLocal,JitterDDP	percentile range 0-1
	F0final, ShimmerLocal	percentile 1.0/99.0, upleveltime 75/90

subset search algorithm are as follows.

Given a set of features $A = \{a_1, a_2, ..., a_d\}$, consider each feature of A as a candidate feature subset. The candidate features are evaluated, and an optimal subset is obtained, which is assumed as $\{a_5\}$. Then, $A' = \{a_5\}$ was chosen as the subset. Subsequently, a new feature from A is added to A' to form a candidate subset containing two features, and the selection criterion is that the novel A' is optimal and better than the original A'. Supposing that the novel feature is $\{a_7\}$, the new subset will be $A' = \{a_5, a_7\}$. These steps are repeated until the optimal feature candidate subset A' is inferior to the previous A' in the $(k+1)th$ round, upon which the search is stopped. The feature subset obtained by the kth round is the optimal feature subset. This strategy of selecting features is referred to as the sequential forward feature selection (SFS) [7]. Other subset search algorithms have similar ideas. Sequential backward selection (SBS) starts from a complete feature set A, removes one redundant feature at a time, and finally obtains the optimal A'. Sequential forward float feature selection (SFFS) [33] starts from the empty set; a subset $X = \{x_i | x_i \in A\}$ is selected from the unselected features in each round to optimize the evaluation function after adding subset X, and then a subset $X' = \{x'_i | x'_i \in X\}$ is selected from the selected features to optimize the evaluation function after eliminating subset X'.

The feature weighting algorithm assigns weights to the features individually by correlation analysis and ranks them based on the value of the weights. When the correlation weight of the feature is larger than a set threshold, it is selected. There are numerous methods for correlation evaluation, such as Euclidean distance, Pearson correlation, and mutual information [18].

The Euclidean distance calculates the distance between two features.

$$E(a_i, a_j) = \sqrt{a_i^2 - a_j^2} \tag{1}$$

The Pearson correlation measures whether two features are on the same line.

$$P(a_i, a_j) = \frac{cov(a_i, a_j)}{\sigma_i \sigma_j} \tag{2}$$

Mutual information indicates whether there is a relationship between two features.

$$I(a_i, a_j) = \sum_{a_i} \sum_{a_j} P(a_i, a_j) \log \frac{P(a_i, a_j)}{P(a_i) P(a_j)} \tag{3}$$

where a_i and a_j are two features of the feature set A. $cov(a_i, a_j)$ is the covariance of a_i and a_j. σ_i and σ_j are the standard deviations of a_i and a_j, respectively. In mutual information, $P(a_i, a_j)$ is the joint distribution, and $P(a_i), P(a_j)$ are the marginal distributions for a_i and a_j, respectively.

ReliefF [23] is a well-known feature selection algorithm based on feature weighting. Its main idea is to use Euclidean distance as the relevant index, then weigh the features according to the degree of distinction between different types of samples and the nearest ones. Fast correlation-based filter (FCBF) [45] is a selection algorithm based on the symmetrical uncertainty (SU) value. The algorithm calculates the SU values of each feature and category to select features higher than the threshold value, as well as the SU values between the selected features and other features to further select a better feature set.

Several feature selection algorithms mentioned above have been applied in SER. This study mainly uses different feature selection methods to compare and analyze the algorithm that is more suitable for speech features, and the selected features are fed into the proposed feature transformation model for classification.

V. Emotion Category oriented Feature weighted

The research on speech features shows that each speech feature exhibits different amplitudes and trends under different emotional states, meaning that each feature has a unique ability to express different emotions. This study proposes an ECFW method to explore the potential expression ability of speech features. The features are assigned different weights in different emotion types so that they can perform to their maximum capacity to recognize the emotion state. The framework of this method is shown in **Fig.2**.

[scale=0.58]images/weights.jpg

Fig. 2. ECFW model architecture

First, a set of training samples $X = \{X_1, X_2, ..., X_C\}$ and their corresponding training labels $Y = \{y_1, y_2, ..., y_C\}$ are specified, where C represents the number of emotion types.

Second, the feature selection algorithms are utilized to select the best features from X, and the selected features $X' = \{X'_1, X'_2, ...X'_i, ...X'_C\}$ are in d dimensions. Here, $X'_i = \{x'_{i1}, x'_{i2}, ..., x'_{in_i}\}, i \in 1, 2, ..., C$, where n_i is the number of training samples from the ith emotion. The weights $W = \{W_1, W_2, ..., W_i, ...W_C\}$ can be calculated for each emotion category according to the following rules.

1) Mean Value: The mean value reflects the trend of data concentration and can be intuitively compared with other samples.

$$\mu = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{n_i} x'_{ij}, \text{ where } N = \sum_{i=1}^{C} n_i \tag{4}$$

$$\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x'_{ij} \tag{5}$$

where $\mu = \{\mu_1, \mu_2, ..., \mu_d\}$ and $\mu_i = \{\mu_{i1}, \mu_{i2}, ..., \mu_{id}\}$ refer to the inter-class mean and intra-class mean in the d-dimension, respectively. C represents the number of emotion types.

2) Variance: Variance describes the degree of dispersion of a random variable, which reflects the fluctuation of the random variable near its expected value.

$$V_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x'_{ij} - \mu_i)^2 \tag{6}$$

where μ_i refers to the intra-class mean calculated by Eq.5 and V_i is in d dimensions.

3) Feature Selection Score: We rank the features by scoring their correlations with the feature selection algorithm. The features with higher scores can obtain higher feature weights than those with lower scores.

$$Rank_i = \frac{d-k}{d}, \text{where } k = 1, 2, ..., d \tag{7}$$

where d denotes the feature dimensions. Finally, the weight calculation formula is as follows.

$$w_{ik} = \frac{\mu_{ik}}{\mu_k} Rank_i V_{ik}, \text{where } k = 1, 2, ..., d \tag{8}$$

Therefore, through the above methods, the weights corresponding to each emotion are obtained. Hence, for the ith emotion, the corresponding weight is $W_i = \{w_{i1}, w_{i2}...w_{id}\}$.

The following step comprises training and testing of the model M. The training part of our SER method can be described as **Algorithm V**. [!htbp] Training part of our proposed model [1] The training sample $\{X_1, X_2 \cdots X_C\}$ and the utterance label $\{y_1, y_2 \cdots y_C\}$, where C denotes the number of emotions. The completed trained deep model M Select speech features by the selection algorithm obtained from the previous experiment, and the selected features $X'_i = \{x'_{i1}, x'_{i2} \cdots x'_{in_i}\}, i = 1, 2, ..., C$ are obtained, where x'_{in_i} is in d dimensions; Calculate the inter-class mean μ and intra-class mean μ_i by Eq.4 and Eq.5 for X'_i; Calculate the variance V_i and feature selection score $Rank_i$ by Eq.6 and Eq.7 for X'_i; Calculate and save the weights

$W_i = \{w_{i1}, w_{i2} \cdots w_{id}\}$ according Eq.8 for features $X_i^{'}$ of each emotion, where d is the feature dimension after feature selection; Normalize $X_i^{'}$ by zero-mean normalization, and multiply the weight value W_i to obtain a new feature representation $W_i X_i^{'}$; Train the deep model M by the new features; M

Following the **Algorithm V**, a well-trained model M is obtained. In the testing phase, we must pay attention to the fact that the emotion category of the test sample is unknown, such that we multiply the weight W_i of each emotion i for the test sample $x^{'}$ after feature selection, and then utilize the new utterance $W_i x^{'}$ to obtain i categories of emotion probabilities through the model M. The output probabilities of the emotional categories are obtained by the last layer of M, which is usually the non-linear function softmax.

$$softmax(y_i) = \frac{e^{z_i}}{\sum_{i=1}^{C} e^{z_i}} \tag{9}$$

where z_i is the ith output of the softmax layer, and y_i denotes the output probability of the emotion i. The final emotional state is determined based on the maximum classification probability of the corresponding emotional i, obtained by multiplying with the weight W_i. The details of the testing part of our SER method are described in **Algorithm V**.

[!htbp] Testing part of our proposed model [1] A test sample x from the test set. The weight value $W_i, i = 1, 2, ..., C$, where C denotes the emotion category. The trained deep model M. The emotion label l of this utterance. Use the feature selection algorithm on x to form a new sample x'; Normalize x' by zero-mean normalization, and multiply each weight value W_i to construct c utterance $\{x'W_1, x'W_2 \cdots x'W_C\}$; i = 1 : C Obtain the probability of emotion classification $\{p_{i1}, p_{i2}, \cdots, p_{iC}\}$ by model M for utterance $x'W_i$, where p_{iC} denotes the probability of the emotional state C; Lets $P = P(:, p_{ii})$; Choose $l = \max_i(P_{ii})$ as the result label l;

l

VI. Experimental Results and Discussion

This section presents details about the database, the experimental setup, the conducted experiments, and the analysis of the results.

A. Data Set

In this study, we use three different speech emotion databases, namely, the Berlin dataset [16], IEMOCAP dataset [9], and MASC dataset [43] to conduct our experiments.

The Berlin emotional database is a German database recorded by the Berlin Polytechnic University. Ten professional actors (five men and five women) in a noise-and-echo-less soundproof room uttered ten identical German short sentences with seven kinds of emotions. The number of speech files in the seven emotion categories is anger (127), anxiety fear (69), boredom (81), disgust (46), happiness (71), neutral (79), and sadness (62). We selected four emotions (happy, angry, sadness, neutral) and included 339 sentences as the experimental data to compare with other datasets with the same emotion.

The IEMOCAP database is an English database recorded by the University of Southern California, which is composed of five sessions from ten professional speakers (five men and five women) in a professional studio. Audio and video content of 12 hours are recorded, including improvisation and pre-script. The annotation of the database is decided by three experts through a majority of votes, which means utterances that were labeled differently by all three annotators were discarded in this study. For all ten emotions, we select five for our experiments: angry (1103), excitement (1040), happy (595), neutral (1708), and sad (1084). Because the expressions of happiness and excitement are similar, we combine these two emotions into one category. Thus, there are a total of 5530 utterances and four emotions for the experiments.

The MASC database is a Chinese database that is the only mandarin emotional prosody speech database published by the International Linguistic Data Union. The database consists of 68 speakers (23 women and 45 men). These speakers are proficient in Mandarin Chinese and can express their feelings clearly. Each recording of the voice database requires the speaker to express the same sentence in five emotions, namely, anger, happiness, painful, sadness, and neutral. The phonetic content includes phrases, short sentences, and paragraphs. The lengths of phrases, short sentences, and paragraphs are 1–2 s, 2–12 s, and 30–60 s, respectively. The database contains 25,636 speech utterances, 5000 phrases, 20400 short sentences, and 136 paragraphs. The sample rate of the corpus is 8000 Hz. In this study, we also use four emotions (happy, angry, sadness, and neutral) from the data set, and a total of 16319 sentences for experiment.

The experiments are conducted in an eight-fold leave-one-speaker-out (LOSO) cross-validation scheme [26]. Therefore, for Berlin and IEMOCAP databases, the first four sessions from eight speakers are selected as training data, and the one from other two speakers are used to cross-validate the hyper parameters of the models, and the last one is used as test data. For the MASC database, the first 80% data, that is, 55 speakers, are selected as training data. As for the remaining 13 speakers, we select the first seven as validation data and the last six as test data.

B. Experiments Setup

The speech emotion is classified through the deep learning model, and the proposed method is tested on Chinese corpus of MASC, English corpus of IEMOCAP, and German corpus of the Berlin corpus. The opensmile toolkit [15] is used for speech processing and analysis. For each voice audio, the first 2 s of speech audio is selected as audio signal input, and the speech audio lasting below 2 s is complemented by adding 0. The speech audio is transformed into segments by using a 25 ms Hamming window with a stride of 10 ms. The neural networks and training algorithms are implemented in Matlab and Python. The experiments are performed on the computer of the 64-bit windows Sever 2012 R2 standard system; the CPU is the dual-core Intel CORE E5, the clock frequency of the processor is 2.1 GHz, and running memory is 128 G. To measure the performance of the systems, the weighted accuracy (WA, accuracy) and unweighted accuracy (UA, average recall over different emotion categories) are reported on those three corpora. Furthermore, MASC is a balanced database, its WA and UA are the same.

C. Matching between features and models

To obtain the relationship between the feature and the model, some commonly used speech features are extracted and matched with several deep models to find an optimal combination of the feature and model. Speech features used in this experiment include two types, LLDs and HSFs.

LLDs in this experiment include LPCC with 16 coefficients, 13 MFCC with its first and second derivatives, the speech spectrogram, and filter banks with 40 coefficients. Three typical deep learning models, CNN, LSTM, and CNN+LSTM, are used to learn the LLDs features and verify emotion classes. CNN consists of two convolutional layers with 32 and 64 kernels per layer, one max pooling layer after each convolutional layer, and two fully connected layers with 1024 nodes. Each convolutional layer has the same kernel size of 3×3, the same strides 2×2, and the same padding. The LSTM model consists of two LSTM layers, containing 512 and 256 nodes, respectively. The CNN+LSTM model combines two convolutional layers, two max pooling layers, and two LSTM layers. The detailed network settings are the same as mentioned above. A dropout layer with a retention possibility P = 0.5 is used after each LSTM and fully connected layer. BatchNorm is used after each convolutional and fully connected layer.

Moreover, the t test [34] statistic comparison is used to measure the similarity between CNN+LSTM and other compared classifiers. We use the p-value to represent significant dif-

ferences. The p-value is obtained by the two compared classifiers on the testing sets. A small p-value indicates a significant difference between the two compared classifiers. Generally, the threshold of the p-value is set to 0.05. When the p-value is less than 0.05, a significant difference between the compared two classifiers in the correct classification is indicated. P-values smaller than 10^{-4} are set to 0.0001. The experiment results are shown in the **Table II**. The asterisk is used to highlight these p-values, which are smaller than 0.05.

TABLE II

RESULT OF LLDS OBTAINED WITH THE DIFFERENT DEEP LEARNING MODELS THROUGH THREE EMOTIONAL DATABASES (MEAN ACCURACY (%) WITH STANDARD DEVIATION AND t TEST, EMO-DB = BERLIN DATABASE, MFCC(13)+Δ^2 = MFCC(13) WITH ITS FIRST AND SECOND DERIVATIVES)

3[3]*Datasets 3-8	3[3]*Features	CNN+LSTM		CNN		LSTM	
		2[1]*WA	2[1]*UA	WA	UA	WA	UA
				p-value	p-value	p-value	p-value
10[2]*Emo-DB	Speech Signal	55.67±1.94	48.21±1.53	51.98±2.55	41.3±2.09	43.41±2.39	41.81±2.21
				0.0500	0.0007*	0.0008*	0.0014*
	MFCC(13)+Δ^2	70.65±1.82	66.3±3.11	68.29±4.08	64.55±2.61	62.63±1.41	54.23±3.08
				0.0432*	0.0475*	0.0011*	0.0001*
	Filter Bank(40)	**73.09±3.58**	**68.79±4.67**	70.57±3.54	66.05±2.55	67.7±2.70	58.57±2.85
				0.0594	0.0627	0.0057*	0.0013*
	LPCC(16)	72.32±2.90	67.25±3.44	54.64±2.22	51.19±2.15	68.5±1.86	59.59±4.12
				0.0009*	0.0018*	0.0042*	0.0041*
	Spectrogram	57.01±2.11	46.45±1.58	63.09±1.38	59.37±1.18	62.64±1.48	56.98±0.92
				0.0013*	0.0001*	0.0024*	0.0002*
10[2]*Iemocap	Speech Signal	47.1±1.27	49.14±1.57	36.03±0.91	33.67±0.91	38.46±1.16	43.25±0.94
				0.0001*	0.0001*	0.0006*	0.0002*
	MFCC(13)+Δ^2	47.8±1.48	49.56±0.54	43.48±1	45.71±0.81	38.48±1.62	39.84±0.97
				0.0013*	0.0021*	0.0031*	0.0001*
	Filter Bank(40)	**51.03±1.13**	**51.75±0.97**	49.75±0.87	51.01±0.88	42.37±1.26	43.35±1.21
				0.1109	0.2889	0.0009*	0.0007*
	LPCC(16)	48.76±1.02	50.92±0.94	43.57±1.64	46.77±1.76	42.69±1.33	45.87±2.36
				0.0007*	0.0032*	0.0001*	0.0041*
	Spectrogram	40.5±1.33	41.69±0.90	39.18±1.56	36.1±3.59	38.89±1.78	41.17±3.55
				0.2337	0.0166*	0.1879	0.7853
10[2]*MASC	Speech Signal	51.26±0.81	51.26±0.81	43.7±0.68	43.33±1.14	43.68±0.54	43.68±0.54
				0.0013*	0.0021*	0.0001*	0.0001*
	MFCC(13)+Δ^2	61.86±1.28	61.86±1.28	56.88±4.59	56.88±4.59	60.63±0.75	60.63±0.75
				0.0697	0.0697	0.1352	0.1352
	Filter Bank(40)	**63.89±0.62**	**63.89±0.62**	53.71±1.10	53.71±1.10	62.9±0.82	62.9±0.82
				0.0001*	0.0001*	0.0887	0.0887
	LPCC(16)	54.17±1.23	54.17±1.23	49.48±0.70	49.48±0.70	55.42±1.23	55.42±1.23
				0.0002*	0.0002*	0.1883	0.1883
	Spectrogram	59.3±1.61	59.3±1.61	55.52±0.78	55.52±0.78	56.07±1.32	56.07±1.32
				0.0029*	0.0029*	0.0145*	0.0145*

Means that the p-value is below 0.05.

From the **Table II**, the best recognition accuracies for the EMO-DB, IEMOCAP, and MASC are 73.09%, 51.03%, and 63.89%, respectively, which are all based on CNN+LSTM model under the feature of Filter Bank(40). **Table II** lists all p-values to show the significant differences

between CNN+LSTM and other compared classifiers. According to the p-value, CNN+LSTM exhibits a significant difference when compared with other deep models on three emotional data sets with various speech features. These results show that for the SER task, the best fit deep learning model to deal with the LLDs features is the combination of CNN and LSTM. Based on CNN+LSTM, the best LLDs feature to be used for classifying the speech emotion is the filter bank. Further, the original speech signal has the lowest rate of emotion recognition in this experiment; hence, it is not suitable for direct application for emotion recognition.

After comparing the matching between LLDs and deep models, we conduct some experiments on the commonly used HSFs. Because the type of HSFs feature is in the form of a vector, we select DNN of different layers as the learning model. 2L-DNN and 3L-DNN consist of two and three fully connected layers, respectively, with 1024 nodes. BatchNorm and Dropout are also used after each layer, and they exhibit the same settings as before. The experimental result is shown in the **Table III**.

TABLE III

HSFs RESULT OBTAINED BY DIFFERENT DEEP LEARNING MODELS THROUGH THREE EMOTIONAL DATABASES (MEAN ACCURACY (%) WITH ITS STANDARD DEVIATION)

2[4]*Architecture	2[4]*Database	Interspeech09		Interspeech'10		Gemaps		eGemaps	
3-10		WA	UA	WA	UA	WA	UA	WA	UA
3[1]*2L-DNN	EMO-DB	**72.48±6.21**	**70.13±5.06**	70.12±5.35	68.33±3.83	67.54±5.28	62.69±5.63	70.14±4.63	66.68±1.84
	IEMOCAP	52.15±0.68	56.24±0.94	55.37±0.71	**58.56±0.41**	56.26±0.9	52.15±1.13	**56.63±0.7**	52.29±0.97
	MASC	62.65±0.38	62.65±0.38	**66.32±0.83**	**66.32±0.83**	62.76±0.93	62.76±0.93	63.01±1.78	63.01±1.78
3[0]*3L-DNN	EMO-DB	**71.61±5.41**	**68.65±3.28**	68.75±4.73	66.3±3.96	65.88±5.32	60.51±6.27	69.85±4.57	66.49±1.86
	IEMOCAP	52.17±0.8	57.46±0.5	**56.28±0.54**	**59.83±0.71**	55.2±1.12	51.83±0.51	55.61±0.84	51.51±0.63
	MASC	61.66±1.17	61.66±1.17	**64.33±0.59**	**64.33±0.59**	62.09±0.77	62.09±0.77	62.83±0.96	62.83±0.96

For EMO-DB, when the feature set is Interspeech'09, and the model is 2L-DNN, the best emotional accuracy 70.13%, and an utterance accuracy of 78.10% is obtained. For IEMOCAP, when the feature set is Interspeech'10, and the model is 3L-DNN, the best emotional accuracy 59.83% is obtained, and when the feature set is eGemaps and the model is 2L-DNN, the best utterance accuracy 56.63% is obtained. For MASC, when the feature set is Interspeech'10, and the model is 2L-DNN, the best emotional accuracy 66.32% and utterance accuracy 66.32% are obtained. The result in **Table III** shows that a shallow network structure (2L-DNN) is sufficient for most data sets when DNN is used for the speech emotion recognition task. Moreover, for

different data sets, the size of data is key to the selection of speech features.

Comparing **Table II** with **Table III**, under the same dataset, most HSFs feature sets are observed to be more powerful than LLDs in SER. This may be because HSFs carry more feature information than a single LLD. Except for the EMO-DB, the best recognition accuracies of IEMOCAP and MASC are obtained through the combination of DNN and HSFs. Meanwhile for EMO-DB, the best result of DNN with HSFs is only 0.5% lower than that of LLDs with CNN + LSTM. Therefore, the feature set of HSFs is considered as the optimal feature in this study. **Table III** clearly shows that for the DNN network, the most suitable HSFs feature set is Interspeech'10. Although Gemaps and eGemaps have better classification results on IEMOCAP datasets, the UA of the two feature sets is generally lower than the value of WA, which means that the features selected by the two feature sets have better classification performance for specific emotions, but are not suitable for general emotional classification. Further, considering the risk of over-fitting, the Interspeech'09 feature set is recommended for EMO-DB, which has a smaller data size.

D. Comparison of feature selection algorithms

Through the last experiment, we obtain a set of features and models with a high recognition rate. However, the feature dimensions are very high, at 1582 dimensions. To reduce the dimensionality and obtain an optimal matching between feature selection algorithms and features, we choose several feature selection algorithms to compare the results. Some are based on feature weighting algorithms, such as ReliefF, MRMR [32], LaplacianScore [19], and FCBF. The other two are based on the subset search algorithms, such as SFS and SFFS. The feature set we use in this experiment is Interspeech'10, which shows good performance in the last experiment. The 2L-DNN model is chosen as the test model.

The weight-based feature selection algorithm will score each feature based on a certain criterion, to obtain the importance of each feature. Thus, we rank the features by these scores and take out the features according to a certain proportion based on the ranking to observe the classification results. The result is shown in **Table IV**.

Subset search algorithms make greedy choices based on specific evaluation indicators, as shown in **Table V**. Further, the FCBF algorithm not only calculates the correlation score, but also eliminates redundant features; hence, we put the FCBF algorithm into the **Table V**.

TABLE IV

MEAN ACCURACY(%) WITH ITS STANDARD DEVIATION OF TWO-LAYER DNN ON SELECTED FEATURES FOR EACH WEIGHT-BASED FEATURE SELECTION ALGORITHM

2[3]*Method	2[3]*Database	Number of features selected			
3-6		100	200	300	400
3[2]*ReliefF	EMO-DB	55.61±2.84	63.41±2.67	**68.78±1.83**	68.29±1.54
	IEMOCAP	53.16±0.23	57.45±0.95	**58.56±0.62**	57.49±0.67
	MASC	61.19±0.57	63.74±0.4	66.36±0.41	**66.38±0.24**
3[2]*MRMR	EMO-DB	63.41±1.54	65.36±0.98	65.36±1.83	**66.83±2.49**
	IEMOCAP	55.84±0.62	56.48±0.86	**58.22±0.84**	56.75±1.01
	MASC	63.43±0.41	65.26±0.34	65.71±0.26	**65.82±0.53**
3[1]*Laplas	EMO-DB	68.29±2.18	**69.27±1.95**	68.29±1.54	67.31±2.49
	IEMOCAP	54.87±0.92	54.84±1.24	**57.62±0.71**	54.67±1.5
	MASC	61.01±0.32	67.12±0.24	**68.58±0.62**	64.93±0.62

TABLE V

MEAN ACCURACY(%) WITH ITS STANDARD DEVIATION OF TWO-LAYER DNN ON SELECTED FEATURES FOR EACH SUBSET SEARCH FEATURE SELECTION ALGORITHM

2[4]*Method	EMO-DB		IEMOCAP		MASC	
2-7	WA	No.	WA	No.	WA	No.
SFS	81.46±1.95	82	53.43±0.45	211	63.26±0.3	303
SFFS	80.49±1.54	94	53.97±0.75	223	62.89±0.71	241
FCBF	72.19±4.52	58	56.34±0.75	204	64.42±0.29	228

The accuracy of each selection algorithm is listed in **Table IV** and **Table V**. For small sample databases, such as EMO-DB, the subset search algorithm is recommended as the feature selection algorithm. Through the SFS algorithm, the optimal number of features is 82, and the accuracy is 81.46%. When using the weight-based feature selection algorithm, the number of excellent features is the top 300 in the ranking. As shown in **Table IV**, for EMO-DB and IEMOCAP datasets, the first 300 features obtained by the ReliefF algorithm yielded 68.78% and 58.56% optimal results, respectively. For MASC datasets, 68.58% is obtained when choosing the first 300 features obtained by the Laplacian Score algorithm.

Fig. 3 shows the trend diagram of ReliefF, MRMR, and Laplas on three databases. Curves are fitted by third-order polynomials. Among them, the abscissa represent the number of features selected according to the ranking of scores, and the ordinate represents the classification accuracy.

Evidently, when the selected features are between 300 and 400, all three databases yield a good classification result. For the Emo-DB, the feature selection effect of Laplas is stable and excellent. For the other two datasets, MRMR is recommended as a feature selection method.

[width=1]images/iemocap.jpg

(a) Accuracy Trend on IEMOCAP
[width=1]images/masc.jpg

(b) Accuracy Trend on MASC
[width=1]images/emodb.jpg

(c) Accuracy Trend on Emo-DB

Fig. 3. Accuracy Trend(%) comparison of ReliefF, MRMR, and Laplas on three databases with different number of features

E. Exploiting the potential information of features

After experimenting on the matching relationship between features and models, a set of features and models with good recognition rate is obtained. According to this combination, we conduct a validation experiment on the proposed method. Based on the first 300 features of each feature selection algorithm obtained from the previous experiment, we multiply the corresponding weights, and train them by a 2L-DNN model, which is used in the above experiments. Moreover, we employ the t test to verify the effectiveness of the proposed method. The results are shown in **Table VI**.

TABLE VI

MEAN ACCURACY (%) WITH ITS STANDARD DEVIATION AND t TEST OF FEATURE WEIGHTING BASED ON FIRST 300 FEATURES OF EACH FEATURE SELECTION ALGORITHM

3[3]*Database	ECFW			Baseline		
2-7	2[1]*RelieF	2[1]*Laplas	2[1]*MRMR	RelieF	Laplas	MRMR
				p-value	p-value	p-value
2[1]*EMO-DB	71.22±0.98	71.71±1.20	72.19±1.95	68.78±1.83	68.29±1.54	65.36±1.83
				0.0462*	0.0081*	0.0086*
2[0]*IEMOCAP	60.83±0.40	60.29±0.34	59.59±0.57	58.56±0.62	57.62±0.71	58.22±0.84
				0.0003*	0.0001*	0.0981
2[1]*MASC	67.83±0.91	69.92±0.83	69.46±0.77	66.36±0.41	65.71±0.26	68.58±0.62
				0.018*	0.0331*	0.0016*

Means that p-value is less than 0.05.

Table VI clearly shows that our system has been improved by 2% to 8% under three different baselines. The best rate achieved in IEMOCAP, EMO-DB, and MASC is equal to 60.83%, 72.19%, and 69.92%, respectively. These improvements on three different databases reflect the good generalization of our system. From the p-values shown in **Table VI**, most are smaller than 0.05, which means the ECFW exhibits a significant difference when compared with the baseline on three data sets using different feature selection method. Thus, the effectiveness of ECFW can also be validated with the t test.

The recognition rate on EMO-DB and IEMOCAP obtained from the proposed method was compared with previous research conducted with the same experiment setup, and the results are tabulated in **Table VII**. **Table VII** moreover indicates that the proposed ECFW obtains a markedly higher recognition accuracy in SER compared with several previous studies.

TABLE VII

COMPARISON OF PROPOSED RESULTS WITH PREVIOUS STUDIES. (ACCURACY (%))

2[2]*Method	2[2]*Features	Accuracy	
3-4		EMODB(SI)	IEMOCAP(SI)
Luo D. et al.(2018)[26]	log-mel spectrogram, Interspeech'16	-	60.35
Cho J. et al.(2018)[10]	Gemaps	-	59.63
Neumann M. et al.(2017)[28]	Log-mel spectrogram	-	56.10
Gangamohan P. et al.(2015)[21]	MFCC	65.60	-
Daneshfar F. et al.(2020)[11]	MFCC, PLPC, PMVDR, Pitch	68.89	-
proposed	Interspeech'10	**72.19**	**60.83**

VII. CONCLUSION

This study proposes a novel ECFW method to recognize speech emotion. The ECFW is designed to explore the features' potential with respect to three aspects. The first is to find the most suitable model that can extract the most valuable information from the features for emotional representation; the second is to employ feature selection methods to select the distinctive emotion-relevant feature subsets; the third is to exploit the emotional information carried by the features and maximize the emotional perception ability of the features by a feature weighting algorithm. Specifically, various speech features, such as the spectrogram, MFCC, and LFCC, are extracted from speech signals, which are matched with the existing deep models such as DNN, CNN, LSTM, and CRNN, respectively. The best combination of features and models suitable for

emotion classification can thus be selected. Subsequently, several feature selection methods, e.g., ReliefF, MRMR, and SFS, are applied to reduce the feature dimensions and select the most discriminative feature subsets for the obtained optimal combination. Finally, ECFW endows each feature under different emotions in the selected feature subsets with the corresponding weights. The proposed ECFW is tested on three benchmark databases (EMO-DB, IEMOCAP, and MASC). The experimental results demonstrate that ECFW has the ability to exploit the potential of features and thus exhibits better speech emotion recognition performance.

However, the generation mechanism of emotion is very complicated. To date, scientists have not been able to unravel its mystery. Even if emotions are labeled manually, such as the data set of IEMOCAP, almost 30% of the data have different annotation opinions determined by three annotators. Thus, the proposed method can improve the recognition of speech emotion to some extent, yet it still cannot meet the needs of practical applications. Moreover, the speech emotion recognition method proposed in this paper is based on the entire utterance. Because emotion has a temporal structure, the entire utterance may contain complex emotions. An exciting future direction of our study is to recognize the compound emotions in an utterance. Furthermore, this work also studies the unique identification of features and uses weights to calculate them. Nevertheless, the corresponding weight must be calculated for each emotion, which implies that the amount of calculation depends on the number of emotion categories. This may increase the computational complexity.

Future studies will focus on studying algorithms inspired by brain mechanisms to simulate the cognitive process of the human brain for emotion recognition to help machines better perceive emotions. Further, the proposed method will be improved and extended to address compound emotion problems, and optimize feature computing and model building to reduce time complexity.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under Grant No. 61806078.

REFERENCES

[1] Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. *Applied Acoustics*, 142:70 – 77, 2018.

[2] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 22(10):1533–1545, 2014.

[3] D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In *Learning from Data - Fifth International Workshop on Artificial Intelligence and Statistics, AISTATS 1995, Key West, Florida, USA, January, 1995. Proceedings.*, pages 199–206, 1995.

[4] JB. Alonso, J. Cabrera, M. Medina, and C. Travieso. New approach in quantification of emotional intensity from the speech signal: emotional temperature. *Expert Systems with Applications*, 42(24):9554–9564, 2015.

[5] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[6] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1322, 1974.

[7] Moataz M. H. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[8] Bachorowski and J. Anne. Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57, 1999.

[9] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[10] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak. Deep neural networks for emotion recognition combining audio and transcripts. In *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*, pages 247–251, 2018.

[11] F. Daneshfar and S. J. Kabudian. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimedia Tools Appl.*, 79(1-2):1261–1289, 2020.

[12] S. Demircan and H. Kahramanli. Feature extraction from speech data for emotion recognition. *J. Adv. Comput. Netw*, 2(1):28–30, 2014.

[13] M. S. Erik, J. S. Jeffrey, and E. K. Youngmoo. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012*, pages 325–330, 2012.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affective Computing*, 7(2):190–202, 2016.

[15] F. Eyben, M. Wöllmer, and B. W. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1459–1462, 2010.

[16] B. Felix, P. Astrid, M. Rolfes, Walter F. Sendlmeier, and W. Benjamin. A database of german emotional speech. In *INTERSPEECH 2005 - 9th Annual Conference of the International Speech Communication Association, Lisbon, Portugal, September 4-8, 2005*, pages 1517–1520, 2005.

[17] Frick and W. Robert. Communicating emotion: The role of prosodic features. *Psychological Vulletin*, 97(3):412–429, 1985.

[18] A. Hacine-Gharbi and P. Ravier. On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition. *Journal of King Saud University - Computer and Information Sciences*, 2019.

[19] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*

18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], pages 507–514, 2005.

[20] K. Huang, C. Wu, T. Yang, M. Su, and J. Chou. Speech emotion recognition using autoencoder bottleneck features and lstm. In *2016 International Conference on Orange Technologies (ICOT)*, pages 1–4, 2016.

[21] S. R. Kadiri, P. Gangamohan, Suryakanth V. Gangashetty, and B. Yegnanarayana. Analysis of excitation source features of speech for emotion recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1324–1328, 2015.

[22] S. Kakouros and O. Räsänen. 3pro - an unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, 82:67–84, 2016.

[23] Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94, European Conference on Machine Learning, Catania, Italy, April 6-8, 1994, Proceedings*, pages 171–182, 1994.

[24] Y. Lecun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[25] J. Lin, C. Wu, and W. Wei. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 14(1):142–156, 2012.

[26] D. Luo, Y. Zou, and D. Huang. Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*, pages 152–156, 2018.

[27] K. Mannepalli, P. N. Sastry, and M. Suman. Emotion recognition in speech signals using optimization based multi-svnn classifier. *Journal of King Saud University - Computer and Information Sciences*, 2018.

[28] N. Michael and T. V. Ngoc. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 1263–1267, 2017.

[29] A. Milton and S. T. Selvi. Four-stage feature selection to recognize emotion from speech signals. *I. J. Speech Technology*, 18(4):505–520, 2015.

[30] S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 2227–2231, 2017.

[31] A. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[32] H. Peng, F. Long, and C. H. Q. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[33] P. Pudil, F. J. Ferri, J. Novovicová, and J. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. In *12th IAPR International Conference on Pattern Recognition, Conference B: Patern Recognition and Neural Networks, ICPR 1994, Jerusalem, Israel, 9-13 October, 1994, Volume 2*, pages 279–283, 1994.

[34] Philip H. Ramsey. Nonparametric statistical methods. *Technometrics*, 42(2):217–218, 2000.

[35] K. S. Rao, S. G. Koolagudi, and R. R. Vempada. Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, 16(2):143–160, 2013.

[36] B. W. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 312–315, 2009.

[37] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan. The INTER-

SPEECH 2010 paralinguistic challenge. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 2794–2797, 2010.

[38] Y. Sun, G. Wen, and J. Wang. Weighted spectral features based on local hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, 18:80–90, 2015.

[39] D. Tang, J. Zeng, and M. Li. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*, pages 162–166, 2018.

[40] E. Tzinis and A. Potamianos. Segment-based speech emotion recognition using recurrent neural networks. In *Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, October 23-26, 2017*, pages 190–195. IEEE Computer Society, 2017.

[41] V. V. R. Vegesna, K. Gurugubelli, and A. Vuppala. Prosody modification for speech recognition in emotionally mismatched conditions. *International Journal of Speech Technology*, 21:521–532, 2018.

[42] K. Wang, N. An, B. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6:69–75, 2015.

[43] T. Wu, Y. Yang, Z. Wu, and D. Li. MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In *Odyssey 2006, The Speaker and Language Recognition Workshop, San Juan, Puerto Rico, 28-30 June 2006*, pages 1–5, 2006.

[44] C. K. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat. Hybrid bbo-pso and higher order spectral features for emotion and stress recognition from natural speech. *Applied Soft Computing*, 56:217–232, 2017.

[45] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 856–863, 2003.

[46] S. Zhang, S. Zhang, T. Huang, and W. Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimedia*, 20(6):1576–1590, 2018.

[47] J. Zhao, X. Mao, and L. Chen. Speech emotion recognition using deep 1d & 2d CNN LSTM networks. *Biomed. Signal Proc. and Control*, 47:312–323, 2019.

[48] L. Zhao, C. Jiang, C. Zou, and Y. Zhen. A study on emotional feature analysis and recognition in speech. *Journal of China Institute of Communications*, 1(4):418–420, 2004.

[49] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li. Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*, pages 272–276, 2018.

[50] W. Q. Zheng, J. S. Yu, and Y. X. Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, pages 827–831, 2015.