Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 230-260
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

Survey on Intelligent Speech Emotion Recognition

Bing Du*, Qingnan Gao, Huansheng Ning
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
*Corresponding Author.

Abstract:

Speech emotion recognition is the most representative branch of affective computing, which is of great significance for achieving harmonious human-computer interaction. This paper systematically surveys intelligent speech emotion recognition algorithms, including speech emotion datasets, classical feature extraction and dimensionality reduction, and various leading emotion classification models. By enumerating the excellent works of speech emotion recognition in the past 20 years, the pros and cons of these works are compared and analyzed. Finally, we summarized the main problems of speech emotion recognition and the future directions, in order to promote human-computer interaction to a more humane stage.

Keywords: *Affective computing, Speech emotion recognition, Classification, Human-robot interaction.*

I. INTRODUCTION

Artificial intelligence has captured the attention of the media and the imagination of the public in computer science, due to the rapid renovation of chips and the boost of novel deep neural networks. Intelligent robots contain the most advanced intelligent technologies and algorithms to demonstrate artificial intelligence. Among intelligent robots, service robot is, by any means, the fastest growing kind of the robot’s industry, because they are capable of doing dangerous and sophisticated tasks due to their excellent mechanical capabilities. Therefore, service robots have achieved extraordinary success in smart medical/health care and home services [1]. However, intelligent robots have few emotions, resulting in the interaction between humans and robots not as smooth and natural as that between humans [2]. As such, many researchers contributed to exploring emotional robots that have the ability to correctly understand human emotions and reactions.

Emotional robots, one branch of artificial psychology, were first proposed by Professor Picard of the MIT Media Lab in the 1990s. After that, various emotion expressions, such as audio signals, facial expressions, and body postures are used to automatically recognize emotions. This has gradually developed into a branch of affective computing, which has experienced fast growth over the last decade [3]. However, human emotions are subtle and intangible. Even humans may not be able to distinguish and understand the emotions conveyed by others through his/her actions, which is more difficult for robots.

Two emotional theories, categorical emotions and dimensional emotions, were proposed to estimate human emotions and have been applied to human-computer interaction (HRI) emotion recognition. Categorical emotion theory divides human emotions into seven discrete categories: happiness, anger, disgust, fear, sadness, surprise and neutrality [4], while dimensional emotion theory believes that emotions are continuous and uses valence-arousal space to describe the emotion. Valence represents a measure of positive and negative emotions, and arousal is the level of emotional activation [5], which can model many comprehensive and subtle emotions [6], as shown in Fig 1.

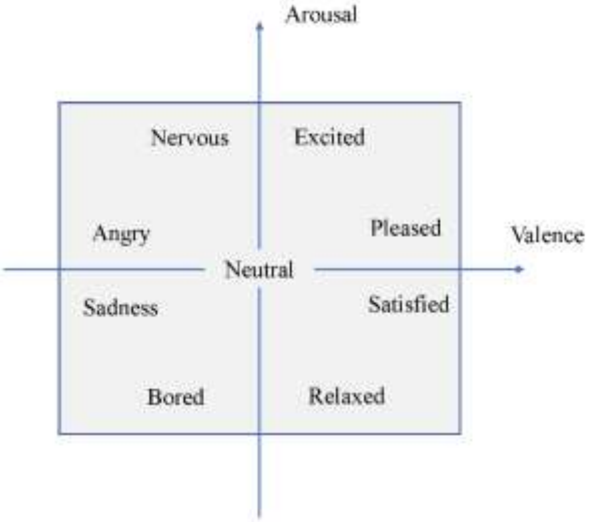

Fig 1: Dimensional emotion theory

Some researchers focus on facial emotion recognition [7-9], while others pay more attention to speech emotion recognition [2, 10-12]. Apart from facial and speech emotion recognition, body gesture emotion recognition is also well developed. Keshari et al. [13] utilized feature-level fusion through combining the extracted features from facial expression and upper body gesture to achieve high emotion recognition accuracy. Other researchers tried to utilize comprehensive multi-modal emotion data such as facial images along with audio streams to recognize emotion. Nguyen et al. [14] utilized two dimensional convolutional autoencoders to produce representative visual features and auditory features, then integrated visual and auditory features into multi-modal features which contain temporal and contextual information by using long short-term memory (LSTM). However, emotion recognition based on facial expressions, audio signals, and body postures can be misleading and deceptive, because these emotional expression channels are an external manifestation of internal emotions and the real emotional state can be deliberately hidden. Therefore, current research pays more attention to physiological signals, such as functional magnetic resonance imaging (fMRI), electroencephalogram (EEG) and galvanic skin response (GSR) [15]. In addition, the micro-expression, which is considered to be the subject’s unconscious emotional expression, may accurately reflect human emotions due to its unconsciousness.

This paper pays more attention on the speech emotion recognition, that is, to infer the speaker's emotional state from their speech. Speech is the most direct channels of interaction either between humans or between humans and robots. People are always able to perceive other's emotional variations through audio signals [16]. An important direction of affective computing is to enable intelligent robots to automatically recognize human emotional states from audio signals [17]. As mentioned earlier, speech emotion recognition can be used for emotional intelligent robots [18], and it can also be used for lie detection and auxiliary treatment [19]. In terms of safe driving, real-time detection of the driver's emotional state through speech emotion recognition can greatly reduce traffic accidents. In online teaching, the emotional state of students can be discovered in time to find out whether the students can understand the knowledge, so as to establish a benign interaction between teachers and students [20]. In order to build a robust and efficient speech emotion recognition system, it is necessary to balance the computational complexity and recognition accuracy. The researchers showed more interests in dimensionality reduction algorithms, extracting compact and representative acoustic features in noisy high-dimensional audio data [21], and an interpretable optimization model for end-to-end learning architecture of speech emotion recognition [22]. Most speech emotion recognition are in a superficial level, which leads to the fact that if a person deliberately hides his inner emotional state, the recognition system is likely to get wrong results. Sometimes, even if a person does not conceal his emotions, there will be a situation where the person is extremely excited but what he said is a bit negative to others, which will lead to the misjudgment of the recognition system.

The rest paper is organized as follows. Section II introduces the speech emotion recognition process, including four fundamental components: emotion datasets, feature categories, feature extraction, and classification algorithms. Section III investigates and analyzes the models and related work. Subjective and objective factors influencing the speech emotion recognition are elaborated in Section IV. Section V derives the existing problems and future directions. Section VI summaries the survey.

II. SPEECH EMOTION RECOGNITION PROCESS

It needs three steps to recognize emotions through audio signals with traditional machine learning. The first step is to select a suitable emotion dataset related to the audio signals, because emotion datasets will affect the recognition accuracy significantly. The second step is to select discriminative and representative speech features, known as feature selection and feature dimensionality reduction. Some feature selection tools can be utilized to extract representative features in order to reduce computation complexity and the redundancy between features. Dimensionality reduction transforms features into a lower dimension. After being processed by dimensionality reduction, speech features have no concrete acoustic meaning, which is obviously different from feature selection. Many approaches can be applicable to feature dimensionality reduction, such as Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS), Isometric Feature Mapping (ISOMAP), Locally Linear Embedding

(LLE), and autoencoder. Feature selection and dimensionality reduction are powerful tools to deal with the curse of dimensionality, which indicates that performance of the classifier becomes worse when the number of features exceeds certain threshold. After these preparations, a representative and compact feature subset is formed. The last essential step is to choose appropriate classifiers to recognize emotional states based on this feature subset. Fig 2 is the emotion recognition process described above. Deep learning, also called feature learning, has injected new vitality into emotion recognition. By establishing an end-to-end learning model, deep learning based methods avoid the defects of traditional machine learning, such as huge feature engineering and high redundancy among selected features by humans.

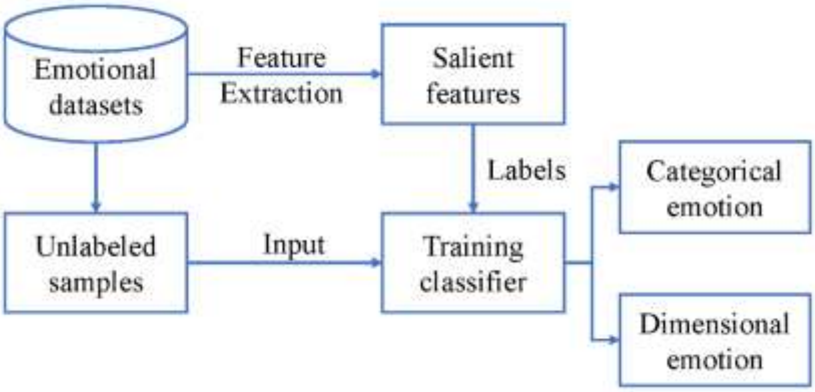

Fig 2: Speech emotion recognition process based on machine learning

2.1 Speech Emotion Recognition Datasets

Strong positive correlations exist between the speech emotion dataset and the speech emotion recognition system. Three major types of speech emotion datasets gained much attentions: acted type, induced type and natural type. The acted emotion datasets may exaggerate emotions resulting in poor generalization performance in the real environment. The induced emotion datasets construct artificial scenes which induce subjects to produce specific emotions spontaneously. Natural emotion datasets collect fragments of real scenes, which may involve issues associated with privacy, thus they cannot be extensively used. Acted emotion datasets include Chinese Emotional Corpus and Self-collected Chinese Speech Datasets (CAESD) [23] and Berlin Emotional Speech database (EmoDB) [24]. Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP) belongs to the category of induced emotion dataset. Natural emotion dataset contains Acted Facial Expressions in the Wild (AFEW) and FAU AIBO [25]. EmoDB dataset is relatively small, while CAESD and IEMOCAP are much larger amongst these emotion datasets [26]. Some of the most widely used speech emotion datasets are as follows:

Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

2.1.1 Berlin emotional speech dataset (EmoDB)

EmoDB is a German acted dataset which is created by the Institute of Communication Science at the Technical University of Berlin [24]. This dataset is widely accepted and applied extensively in speech emotion recognition research. EmoDB is a relatively small dataset and 535 emotional utterances are available. Obvious differences are in the number of different emotional utterances.

2.1.2 Chinese audio emotional speech dataset (CAESD)

The language of the dataset is Mandarin and is widely used in Chinese speech emotion recognition research. The speech data are collected and created by National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences [27]. 300 utterances expressed in six emotions, including neutral, surprise, fear, angry, sad, happy were collected. Four Chinese native speakers consisting of two men and two women recorded 7200 emotional utterances.

2.1.3 Danish emotional speech dataset (DES)

Danish Emotional Speech (DES) is well annotated and extensively used. All utterances are equally separated for each gender. And it contains five emotions including neutral, anger, happiness, sadness, and surprise [28]. Twenty judges participated in emotion annotation.

2.1.4 Remote collaborative and affective interactions (RECOLA)

46 multi-modal recordings involving audio, video, and ECG and EDA were collected in RECOLA [29] with French. RECOLA reflects natural and spontaneous behaviors. Six annotators labeled emotions according to arousal and valence.

2.1.5 Interactive emotional dyadic motion capture dataset (IEMOCAP)

IEMOCAP is established by Speech Analysis and Interpretation Laboratory at the University of Southern California [30]. Five man and five women participated in the recording and the length of recording is approximately 12 hours. The utterances of angry, neutral and sad constitute the majority, and IEMOCAP is thus unbalanced. It is more reasonable to use criterion such as unweighted average recall rate to evaluate the dataset. Both categorical labels and dimensional labels are annotated.

2.1.6 FAU AIBO

FAU AIBO is released through INTERSPEECH 2009 Emotion Challenge. It is a spontaneous emotion dataset based on the interaction between 51 children and a robot named Aibo [31]. The more detailed information about speech emotion datasets is summarized in TABLE I.

TABLE I. Seven popular speech emotion datasets

Dataset	Language	Type	Size	Label
Belfast	English	Acted	40 utterances	Anger, Sadness, Happy, Fear, Neutral
Chinese audio emotional speech dataset	Chinese	Acted	400 utterances	Neutral, Angry, Fear, Happy, Sad, Surprise
DES	Danish	Acted	180 utterances	Anger, Joy, Sadness, Surprise, Neutral
EmoDB	German	Acted	535 utterances	Neutral, Anger, Fear, Joy, Sadness, Disgust, Boredom
FAU AIBO	German	Natural	9.2h	Anger, Emphatic, Neutral, Positive, Rest
IEMOCAP	Multilingual	Induced	12h	Arousal, Valence
RECOLA	French	Natrual	9.5h	Arousal, Valence

2.2 Various Speech Acoustic Features

Speech features and emotional states are closely connected, which plays a decisive role in constructing a robust speech emotion recognition system, and hence it is quite necessary to analyze acoustic features of speech [26]. For instance, happiness is usually accompanied by faster speech rate, while lower intensity and slower speech rate often imply sadness [25, 32]. Acoustic features are usually divided into three categories: prosodic features, voice quality features and spectral features [25]. Fundamental frequency (F_0) and intensity are common prosodic features. Voice quality features include jitter, shimmer, pitch, formant, and bandwidth. Spectral features include linear spectrum and cepstral spectrum. Linear spectral features include linear predictor coefficients (LPC) [33] and one-sided auto-correlation linear predictor coefficients (OSALPC) [34]. Cepstral spectral feature indicates mel-frequency cepstral coefficients (MFCC) [35] and linear predictor cepstral coefficients (LPCC) [36]. Cepstrum is the spectrum processed by logarithm and discrete cosine transform (DCT). MFCC is widely utilized because 12 coefficients of MFCC can be highly representative for speech emotions.

Most of acoustic features are low-level descriptors (LLDs), such as pitch, intensity, and formant [37]. These LLDs usually have plenty of redundancy, which is not helpful to improve the accuracy of emotion recognition. Researchers prefer to use high-level statistical functions (HLFs), such as mean and variance, to further process these LLDs to make features more discriminative for emotions [25, 38]. This does not mean that HLFs are absolutely better than LLDs, although some studies have shown that the features obtained by HLFs achieve higher emotion recognition accuracy with some specific emotion datasets. HLFs' intrinsic defect is that HLFs focuses on the global features at the utterance level and ignores the temporal information contained in the local features of speech signals [39].

In early stage, many emotion recognition systems mainly used spectral features and prosodic features since people thought these features could describe emotion well. Zhou et al. [40] combined prosodic features and spectral features to conduct emotion recognition experiments. They found that formants were less affected by emotional fluctuation, so they discarded this combined acoustic feature. Tao et al. [25] used 2276 acoustic features, including prosodic features, voice quality features and spectral features, to conduct comparative analysis experiments. They found that acted emotion datasets are sensitive to prosodic features and voice quality features, while spectral features are more robust for induced and natural emotion datasets. Their explanation is that prosodic features such as pitch are difficult to extract in wild scenes with noisy.

A detailed description of the common acoustic features is listed at below:

Intensity is the instant sound pressure value and refers to the loudness of sound perceived by human ears.

Pitch related with F_0 is the most commonly used acoustic feature in speech analysis. The movement of vocal cord produces Pitch. In different emotional states, the tension of vocal cords is variant, thus pitch is also different [20].

Jitter mainly reflects the degree of rough for sound and the fundamental frequency change of sound waves between adjacent periods.

Shimmer describes the fluctuation of amplitude between adjacent periods.

Formant refers to some areas where energy is relatively concentrated in the spectrum of sound which reflects the physical characteristics of vocal tract.

Harmonics to noise ratio (HNR) is the ratio of harmonic signals to noise signals in speech, which can effectively reflect glottal closure [10]. The measurement unit of HNR is dB.

Linear predictor coefficients (LPCC) reflect the characteristics of each person’s specific channel. By extracting these coefficients, different emotions can be distinguished [20].

Mel-frequency cepstral coefficients (MFCC) can have speech data parameterized and is widely used in speech recognition and speech emotion recognition [41]. MFCC is highly representative of speech signals. It reflects the local characteristics of speech, while ignores the global and variable information in emotion recognition. In MFCC feature extraction, pre-emphasis is first conducted to compensate the loss of high frequency components in signal and strengthen signal power [42]. Then, the speech signal is divided into short time intervals known as frames. Generally, the proper size of the frame is 20-40 ms. After that, the speech signal will go through a selected windowing function, which aims to smooth the edge of frames and reduce the spectral distortion, such as Hamming window and Hanning window [43]. As for each frame, Fast Fourier Transform (FFT) is conducted to transform the samples from time domain to frequency domain, thus obtaining the corresponding data spectrum. Considering the human auditory mechanism, the linear spectrum is sent to pass through a set of Mel filter banks to be transferred to Mel nonlinear spectrum. Both Mel and Hz are units of sound frequency and the Mel scale conforms more to human auditory perception compared with the Hz scale [44]. Therefore, Mel scale is adopted for analysis. The mapping between Mel and Hz is in Eq. (1).

$$m = 2595log_{10}(1+\frac{f}{700}) \tag{1}$$

Where m denotes Mel scale and f denotes Hz scale. Mel and Hz show linear correlation below 1kHz and logarithmic correlation above 1kHz [45]. Finally, cepstrum analysis is performed, which consists of logarithmic operation and Inverse Fourier Transform. Inverse Fourier Transform is usually implemented by Discrete Cosine Transform (DCT). MFCC feature is thus obtained.

2.3 Feature selection and dimensionality reduction for speech emotion recognition

As one kind of acoustic features, prosodic features have dominated speech emotion recognition for a long time. The reason why researchers are obsessed with prosodic features is that they contain most of the speech information [46]. As time goes by, a trend of combining multiple acoustic features for speech emotion recognition has gained increasing popularity [26]. Shen et al. [20] used prosodic features and spectral features to explore the most representative feature in EmoDB. Their experiment results showed that the recognition accuracy of combining prosodic features with spectral features was higher than using one of them alone. In addition, they also found that spectral features outperformed prosodic features on possessing the edge, which is consistent with Tao et al. in [25]. The prosodic features have large discrepancy in different level of emotional arousal and it is difficult to distinguish emotions with similar arousal values only by prosodic features [47]. Apart from combining with ordinary acoustic features, text features which are generated from automatic speech recognition, so-called lexical features or linguistic features, are also added to the mixed feature set [48]. To some extent, it is reasonable to assume that lexical information contained in the text often reflects the emotional state of speakers, which is beneficial to speech emotion recognition. Jin et al. [49] extracted some LLDs from speech signals, such as intensity, F_0, etc., and took these LLDs with corresponding statistical function features as the acoustic features. At the lexical level, they proposed a new feature representation based on emotion lexicons, named as emotion vector. The bag-of-words (BoW) features with the emotion vector together were called lexical features. Li et al. [2] used text generated from automatic speech recognition for sentiment analysis to determine the emotional polarity of the text, such as positive, negative or neutral, and integrated the obtained information with emotion and prosodic features to conduct emotion recognition, since they thought that the challenge of speech emotion recognition largely attributed to the dependence of emotion expressions on the text content.

Basically, there is no consensus on what is the best acoustic feature and a widely recognized feature subset for emotion recognition has not yet been formed. Indeed, the robust and recognized feature subset should be composed of those features which are representative, discriminative and emotion-related, rather than speaker-related [38]. Researchers used the automatic feature extraction tool named openSMILE [50] to extract a large number of low-level descriptors (LLDs). The simple combination of these features without any further processing does not help to build a highly precise emotion recognition system. Instead of the feature number, feature types should be more emphasized, which can probably depict emotions from different perspectives. Furthermore, the emotion recognition requires short training time and small computational load, thus using thousands of features may run counter to this goal. Therefore, it is necessary to carry out feature selection and dimensionality reduction [51, 52].

For feature selection, heuristic algorithms and random algorithms are usually applied. The former includes Sequential Forward Selection (SFS), Sequential Floating Forward Selection (SFFS) [53], etc., while the latter includes Genetic Algorithms (GA), Simulated Annealing (SA), etc. SFS is essentially one of the greedy algorithms. Normally, the feature subset starts from an empty set. Then, through the evaluation function, the optimal feature is selected and added to the feature subset. SFFS is similar to SFS but at each time adding several features to the feature subset and removing several features to optimize the evaluation function. In [54], the five best features are selected using SFS. Lugger et al. [47] utilized SFFS to select appropriate features from prosodic features and voice quality features. As a typical heuristic algorithm, SA endeavors to find the global optimum rather than the local optimum. GA simulates the natural selection process: first, randomly generate a feature subset; then each of the features in feature subset will be scored by an evaluation function. The higher the score, the more likely it is considered to be strong and reproduces offspring through gene crossover and gene mutation. Repeat this procedure until a good feature subset is generated. Le et al. [12] used GA to choose the most relevant features leading to a good generalization performance for Support Vector Machine (SVM).

At present, most dimensionality reduction algorithms are unsupervised. Dimensionality reduction maps data points in a high-dimensional space to a low dimensional-space, while maintaining original data structure as much as possible and reducing redundancy. Among these dimensionality reduction algorithms, Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Multiple Dimensional Scaling (MDS) are commonly used. LDA tries to keep similar samples as close as possible, while samples with large differences are far away from each other. Notably, LDA is a supervised method. PCA tries to keep the variance between samples to the maximum, while MDS tempts to keep the distance between two samples equal in both high-dimensional space and low-dimensional space. PCA and MDS both need to

Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

perform matrix eigenvalue decomposition. You et al. [55] conducted comparative analysis of PCA, LDA, and the combination of the two in speech emotion recognition. The results confirmed that no one method was the best and each method had a specific application occasion. Zheng et al. [56] utilized PCA to whiten the log-spectrogram and then conducted speech emotion recognition experiment with deep convolutional neural network (DCNN). PCA whitening can improve the recognition accuracy. However, the data structure cannot be truly revealed, because the abovementioned methods assumed that data points are distributed in a linear space [57]. In recent decades, the advent of the manifold learning, such as ISOMAP [58], LLE[59], Lipschitz embedding, etc. partially resolves this problem. In ISOMAP, traditional Euclidean distance is replaced by the geodesic distance to measure the distance between two data points. After obtaining the geodesic distance, MDS is conducted for further processing. You et al. [57] projected 64-dimensional acoustic features into 6-dimensional space with Lipschitz embedding and employed multi-SVM as the classifier. The results demonstrated that significant improvement had been achieved in both speaker-dependent and speak-independent emotion recognition.

A neural network architecture called autoencoder can also be used for dimensionality reduction. An autoencoder is mainly composed of encoder, decoder, and loss function. In order to learn abstract representation of the data, the output of the encoder generally has a lower-dimension than the input of the encoder, thus performing dimensionality reduction. The decoder reconstructs the input and loss function refers to the reconstruction loss. Intensively used autoencoders include sparse autoencoder, denoising autoencoder and convolutional autoencoder.

The fundamental principles of these autoencoders are similar. Take denoising autoencoder for an example. In order to extract more robust features, denoising autoencoder corrupts the training data on purpose and then calculates the Loss function to compare the output z with the original input x, not with the corrupted input, specifically including two stages: pre-training and fine-tuning, as shown in Fig 3.

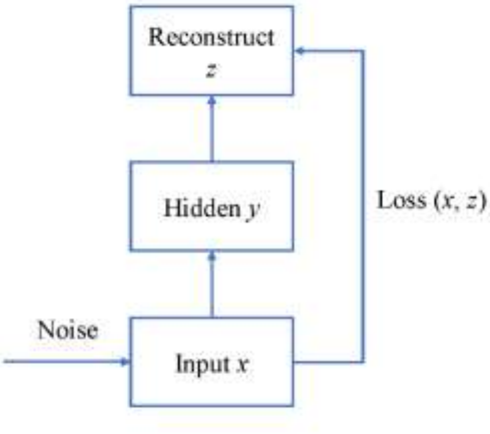

Fig 3: The structure of denoising autoencoder

In pre-training, original input x_o is added with noise n, obtaining the corrupted input x.

$$x = x_o + n \tag{2}$$

Then the encoder compresses the corrupted input x into a latent space representation y through weight matrix W and bias unit b. Afterwards, the decoder reconstructs the input x_o from y.

$$\begin{gathered} y = \sigma(Wx + b) \\ z = \sigma(W'x + b') \end{gathered} \tag{3}$$

$W' = W^T$ means tied weights. σ refers to an activation function such as *Sigmoid* or *Relu*. Reconstructed loss function is the squared error.

$$Loss = \|z - x_o\|^2 \tag{4}$$

After pre-training, labeled data can be utilized in fine-tuning. When these two processes are done, the latent-space representation y can be the input of a well-trained classifiers such as SVM.

Enormous researches recognize speech emotion with autoencoders. The experimental results of Chao et al. [51] confirmed that using denoising autoencoder improved classification accuracy by 7.13% compared with the benchmark in CAESD. Compared to using ordinary acoustic features, Xia et al. [60] found that recognition accuracy improved significantly by using a modified autoencoder in IEMOCAP. In [14], Nguyen et al. utilized two convolutional autoencoders to learn compact and representative features from visual raw data and speech raw data respectively. Their multi-modal emotion recognition achieved state-of-the-art performance in RECOLA. In addition to autoencoder, some other neural networks can also achieve the performance as well as autoencoders. In [6], Chen et al. extracted abstract features from a fifth-convolutional-layer soundnet to carry out their multi-modal emotion recognition. Also, in [38], a ladder network is proposed to learn robust feature representation. Their results showed that although the use of ladder network did not outperform ordinary acoustic features, it achieved a balanced performance in IEMOCAP compared to the benchmark.

2.4 Classification Algorithms in Speech Emotion Recognition

2.4.1 Hidden Markov Model (HMM)

HMM is a canonical generative model, popular in speech emotion recognition. Generally, HMM models short-term acoustic features at frame level to recognize emotion. HMM has three crucial parameters: a state transition probability matrix, an observation probability matrix and an initial state distribution $\lambda = (A, B, \pi)$ [61]. When the observation sequence referring to the frame level features is given, the training stage determines the three parameters A, B and π of the model. As a special form of

EM algorithm, Baum-Welch algorithm is a promising training algorithm, but it can only seek a local optimal solution. For each of the emotional classes, a corresponding HMM is generated in the training stage.

In the speech emotion recognition stage, the observation sequence $O = (o_1, o_2, \dots, o_T)$ extracted from an emotional utterance is input into the HMM model. In order to obtain the prediction of emotions, the forward algorithm is leveraged to calculate the maximum probability of the observation sequence appearing in the model corresponding to each emotion class. The above calculation process is expressed in Eq. 5:

$$P(O|\lambda_e) = \sum_{k=1}^{n} \alpha_T(k), \tag{5}$$

Where λ_e is the parameter vector of the model corresponding to emotional class e; α_T represents the terminal forward variable in the forward algorithm and n is the number of hidden states in HMM.

$$e^* = \arg\max_{1 \le e \le E} P(O|\lambda_e), \tag{6}$$

Where E denotes the number of emotional classes. Eq. (6) shows that the prediction result is identical to the one with the highest probability among these models.

2.4.2 Gaussian Mixture Model (GMM)

In speech emotion recognition, traditional machine learning, such as SVM, mostly employs global features or utterance level features to serve as the input of classifiers. Global features may ignore the temporal information contained in local features, which embodies subtle changes of emotion. GMM is a generative model that can make good use of local features of speech. In this case, the classifier not only needs to accept the input of frame level features, but also cater to emotional utterances of different lengths. However, GMM is computationally intensive and requires lots of labeled samples, so it is difficult to be applied in real-time scenario.

During training, each kind of emotions corresponds to a GMM model respectively [62]. Assuming that there are m kinds of emotions, the GMM model corresponding to emotion j is shown in Eq.7.

$$p_j(x) = \sum_{i=1}^{n} \alpha_i p(\lambda_i, x) \tag{7}$$

$p(\lambda_i, x)$ is a probability density function of Gaussian component λ_i with mean vector μ_i and covariance matrix Σ_i. α_i denotes the weight of Gaussian component λ_i. The probability density function should satisfy the sum of all weights belonging to one kind of emotion equals to 1, e.g. $\sum_{i=1}^{n} \alpha_i = 1$. x often represents the frame-level feature vector. The training of the model is an iterative optimization process based on maximum likelihood estimation (MLE) and expectation maximum (EM) algorithm. Emotion recognition first extracts the feature vector $X = (x_1, \dots, x_K)$ from the emotional utterance, where x_i; $i \in 1,2, \dots, K$ denotes a frame level feature. For the given feature set X, log-likelihood of the emotional utterance under each emotion model j is calculated in Eq.8.

$$r = \arg\max_{1 \le j \le m} \sum_{k=1}^{K} log p_j(x_k) \tag{8}$$

If emotion model r can maximize the log-likelihood of the emotional utterance, the output of emotion recognition is emotion r.

2.4.3 Support Vector Machine (SVM)

SVM is a canonical and discriminant model, which can perfectly adapt to small dataset learning with high classification accuracy and have a solid theoretical foundation. SVM is not sensitive to data dimensions, because it can divide samples with a hyperplane [63]. Training set, $D = \{(x_1, y_1), (x_1, y_1), \dots, (x_n, y_n)\}$, where $y \in \{-1,1\}$, contains n samples. The hyperplane is described in Eq. 9. And the optimization is shown in Eq.10.

$$\omega^T x + b = 0 \tag{9}$$

$$\begin{gathered} \max_{\omega,b} \frac{2}{\|\omega\|} \\ s.t. y_i(\omega^T x_i + b) \ge 1 \end{gathered} \tag{10}$$

ω and $b$in Eq. 9 can be obtained through efficient optimization algorithms. In Eq. (9) and Eq. (10), the samples are sparable linearly. However, for complex acoustic features, linear SVM may fail to separate the data. A kernel function can be employed to deal with the problem by projecting feature vectors to the high-dimensional space. Basically, the kernel function returns the inner product of the two points. Linear kernel (Eq. 11) or Gaussian kernel (Eq. 12) can be chosen.

$$K(x_i, x_j) = {x_i}^T x_j \tag{11}$$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{12}$$

Multi-SVM is more attractive for emotion recognition, like one-to-one and one-to-rest. Assume k kinds of emotions, one-to-one multi-SVM creates $k(k-1)/2$ hyperplanes, while one-to-rest multi-SVM creates $k$hyperplanes.

2.4.4 Convolutional Neural Network (CNN)

As one of the classic representatives of deep neural network, CNN has made great achievements in image processing. Similar to the deep neural network, the back propagation algorithm can be leveraged to train the network and learn parameters. CNN is adept at extracting spatial information, which attributes to the convolution operation. Besides, CNN has two essential characteristics: local connection and weight sharing [64]. Different from ordinary neural network where two adjacent layers are fully connected, CNN only connects part of nodes between adjacent layers, which is called local connection. The weight sharing is to share parameters between two adjacent layers of CNN. Compared with fully connected network, local connection and weight sharing reduce the number of weight as well as lower the risk of over fitting. However, under the premise that the input features are irrelevant to the context, CNN pays little attention to temporal information within emotional utterances.

Convolution layers and pooling layers are essential components of the architecture of CNN and sometimes they may be placed alternately. Full connection layers and the softmax layer are in the end. In the convolution layers, the convolution kernel acts as a local feature extractor. Before training, the parameters are initialized randomly, and then these parameters are trained by the back propagation algorithm. Considering that the speech signal is one-dimensional, the one-dimensional convolution operation is shown in Eq.13.

$$x(n) = \sum_{t=-m}^{m} s(t) * \omega(n-t) \tag{13}$$

$s(n)$ denotes input signal. $\omega(n)$ denotes one dimensional convolution kernel and m denotes the size of it. $x(n)$denotes the result of the convolution. After convolution, a feature map is generated. Then nonlinear activation functions are imposed on the result of the convolution. In the pooling operation, there is no need to learn additional parameters, because this process is a fixed mapping function. And the purpose of pooling operation is to reduce the size of data and prevent the network from over fitting.

2.5.5 Recurrent Neural Network (RNN) and its variants

Recurrent neural networks (RNN) are the state-of-the-art intelligent algorithm for sequential data and are used by Apple's Siri and Google's voice search. Instead of treating the input data isolatedly and irrelevantly, RNN stores the input in its internal memory and captures the temporal information of the sequence, which is preferred for sequential data like time series, speech, financial data, audio, video, weather and much more. RNN shares the weight matrix, applying weights to the current and also to the previous input, as shown in Fig.4. Eq.14 is the formula of hidden states.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
Nov-Dec 2021 Page No. 230-260
Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

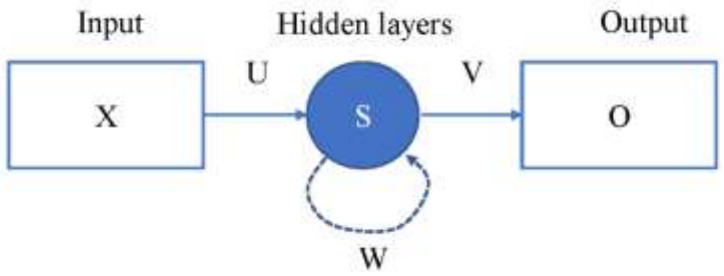

Fig 4: The structure of original RNN

$$s_t = g(U \cdot x_t + W \cdot s_{t-1}) \quad (14)$$

The current output is in Eq. (15).

$$O_t = f(V \cdot s_t) \quad (15)$$

U, V and W are weight matrices and g, f are nonlinear activation functions, e.g., $Sigmoid$ and $Relu$. x_t and O_t are the input and output at time t, respectively. s_t is the state of hidden layers. RNN utilizes x_t and s_{t-1} to predict O_t[65].

RNN suffers from gradient vanishing, resulting in some parameters in the back propagation time (BPTT) algorithm unable to be updated effectively. Long Short-Term Memory (LSTM), a variant of RNN, [66], can deal with the problem of long-term dependence of RNN with cell states and gates. However, the cost of LSTM is the large amount of computation, which is unfavorable for real-time applications.

LSTM introduced the attention mechanism [67] to obtain higher performance when facing extremely long sequential data, which attracts more attentions. The attention mechanism imitates that humans tend to focus more on the target in order to comprehend it. Normally, emotional utterance segments have different levels of emotional saturation, the attention mechanism can make it pay more attention on the segments with higher level of emotional saturation. In LSTM, the attention weight α_t and the context vector c can be obtained by Eq. 16 and Eq. 17, respectively.

$$\alpha_t = \frac{\exp(u^T y_t)}{\sum_{i=1}^{T} \exp(u^T y_t)} \quad (16),$$

$$c = \sum_{t=1}^{T} \alpha_t \, y_t \quad (17)$$

u is the attention vector and y_t is the output at time t.

Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

III. MACHINE LEARNING AND DEEP LEARNING PERFORMANCE COMPARISON

In speech emotion recognition, traditional machine learning classifiers have always used the features manually extracted by humans. Typical algorithms include HMM, GMM, SVM, decision tree, KNN, naive Bayes, etc. Among them, HMM and GMM are generative models, which mainly model the characteristics of the emotional utterance at frame level. Because of this, they can effectively deal with emotional utterances with variable length. But this usually requires inputting a large amount of data into the model, which cannot meet the real-time requirements well.

In [90], HAN et al. used HMM to classify voice emotions. They used a genetic algorithm (GA) to train HMM instead of the traditional Baum-Welch algorithm in order to find the global optimal solution. Considering the use of fixed-size windows in the Fourier Transform, another highlight is that a new method based on wavelet transform replaces Fourier Transform to extract features. Through comparative analysis experiments, they found that the improved HMM is better than the original HMM, and the error rate is significantly reduced. In [91], Chenchah et al. also used HMM for speech emotion recognition. Note that in real life scenarios, voice signals are often mixed with noise. For this reason, three speech enhancement methods were applied in experiments. They incorporated three different noise components into emotional utterances from IEMOCAP. The experimental results showed that the HMM classifier could significantly improve the recognition accuracy only when it adopted speech enhancement method–spectral subtraction, and artificial mixing of training noise. In other cases, the performance did not improve obviously. In [69], Bozkur et al. believed that Line Spectral Frequencies (LSF) can also clearly distinguish speech emotions, compared with MFCC. In their experiments, GMM was used as a classifier and late-fusion strategy, also known as classifier-level fusion, was adopted. The experimental results proved that late-fusion of MFCC and LSF exceeds the recognition accuracy of the model using MFCC alone.

Different from HMM and GMM, SVM is a discriminative model. It is generally believed that SVM has better generalization ability among many machine learning algorithms as shown in Table II. SVM requires fixed feature dimensions. This requirement can be achieved by applying statistical functions to frame-level features. Other methods can also achieve fixed dimensions, such as applying GMM to model each emotional utterance. Each GMM model has the same number of Gaussian components, and their mean vector constitutes the Gaussian supervector. The advantage of SVM is that it does not require a large amount of data, because the construction of the hyperplane is only related to the support vector. As such, SVM is kind of a measurement when evaluating other models [56, 93].

In [40], Zhou et al. designed an SVM architecture based on GMM supervectors, with spectral features. Likewise, Hu et al. [92] compared and analyzed the SVM with GMM, and concluded that SVM is more effective than GMM in gender-dependent systems.

You et al. obtained a 6-D feature with an enhanced Lipschitz embedding of dimensionality reduction and then input this 6-D feature to a linear support vector machine classifier to perform emotion recognition [57]. The results showed that in speaker-independent and speaker-dependent experiments, the classification accuracy can be both improved.

Wang et al. [82] applied improved MFCC features to SVM classifier on EmoDB. They found that only half of the "happiness" state can be recognized, while others are recognized as "anger", because these two emotions are too close along the arousal axis in valence-arousal space.

Some researchers [20, 82] attempted to input multiple features, e.g., prosody and spectral features, to the classifier, resulting in a higher classification accuracy than that with a single feature.

Additionally, a support vector regression based on bag-of-audio-words (BoAW) was provided in [37] to predict emotions on RECOLA dataset. They even discarded some popular CNN and RNN based algorithms. Lee et al [94] constructed a robust decision tree to recognize emotions and on IEMOCAP dataset it can improve the accuracy by 7.44% over SVM. KNN and Bayesian classifier attracted the interest of academia in speech emotion recognition [73, 95] for their consolidated mathematical foundation.

At present, it is agreed that ensemble learning is better than general machine learning in performance. [78] adopted the ensemble learning to deal with the uneven distribution of emotional utterances in FAU AIBO. The recognition accuracy can reach up to 45%. [47] adopted a cascaded-stage classification architecture with prosodic features. They first divided emotions into two activation categories, because in this stage, the prosodic feature can capture the emotion activation accurately, e.g., an average accuracy of 95.5%. And then they further classify the emotions in each of the activation category by the intervals between the emotions. In this cascaded-stage way, an average recognition rate of 74.5% can be achieved.

TABLE II. A summary of dataset features and classifiers in 2000-2015

Literature	Year	dataset	Features	Classifies	Conclusion
[68]	2005	DES	F0, Signal Power, First Four Formant Frequencies, MFCC1, MFCC2 and Five Mel Frequency Sub-band Power	HMM, SVM	The recognition accuracy with HMM was 98.9% for female, 100% for male. The recognition accuracy with SVM was 89.4% for male and 93.6% for female respectively.

[55]	2006	Local Corpus	Prosodic Features, Formant Frequency	SVM	The average recognition rate of male is 83.4% and that of female is 78.7%
[57]	2006	CASIA	48 Prosodic Features and 16 Formant Frequency	Multi-SVM	The proposed system obtained 9%-26% improvement in speaker-independent emotion recognition and 5%-20% improvement in speaker-dependent.
[17]	2007	Local Corpus	Statistic Features	Co-training algorithm with multi-SVM, HMM	The proposed system had a 9.0% improvement on female model and 7.4% on male model.
[40]	2009	CASIA	MFCC, Intensity, F0, Voice Source	SVM	The combination of spectral and prosodic features had higher recognition accuracy than using only one of them.
[69]	2010	EmoDB, FAU AIBO	LSF, MFCC	GMM	The combination of LSF and MFCC outperformed MFCC alone.
[70]	2010	EmoDB	MFCC, MEDC	SVM	The recognition accuracy for gender independent is 93.75%, for male is 94.73%, for female is 100%.
[20]	2011	EmoDB	Intensity, Pitch, LPCC, MFCC, LPCMCC	SVM	The classification accuracy using power and pitch features is 66.02%, the classification accuracy using LPCMCC features is 70.7%, and the classification accuracy using all features is 82.5%.
[71]	2012	EmoDB, Local Corpus	Energy, Pitch, LPCC, MFCC, and MEDC	SVM	The highest accuracy rate on local corpus is 91.3% and EmoDB is 95.1%
[72]	2013	EmoDB, eNTER-FACE	MFCC	DNN-HMM	Among all the models, the DNN-HMM with discriminative pre-training performed best.

[21]	2013	FAU AIBO	MFCC, Intensity, and their first and second temporal derivatives	HMM and DBN	The best unweighted average recall rate (UAR) is 46.36%.
[10]	2014	RECOLA	Intensity, Spectrogram, Pitch, MFCC, HNR, Long-term Average Spectrum, Statistical Features	BayesNet, RBF, SVM	The use of the speech dependent recognition system achieved remark-able improvements.
[73]	2014	EmoDB, CASIA, EESDB	Fourier Parameters, MFCC	SVM, Bayesian	Using FP improves the recognition rates over using MFCC. Combining FP with MFCC, the recognition rates can be further improved.
[52]	2014	CASIA, EmoDB	Speed, DSS, Energy-max, Energy-mean, F_0-max, F4-mean, HNR-SD, F2-SD, F0mean, F4-max, HNR-mean	Decision Tree	For same statements, the recognition accuracy increases from 49.72% to 77.53% with the number of features declining from 384 to 11. For different statements, the recognition accuracy increases from 84.92% to 87.92% with the number of features declining from 33 to 11.
[51]	2014	CASIA, ChongQing Dialect Dataset	32 Low-level Descriptors (LLDs) and 12 Functions	SVM	Denoising autoencoders are used to learn representative features.
[56]	2015	IEMOCAP	Spectrogram	CNN	CNN outperformed the manual featuring on the speech emotion recognition accuracy.
[49]	2015	IEMOCAP	Intensity, F_0, Jitter, Shimmer, Spectral Contours, Emotion Vector Feature (proposed by authors) and Bag-of-Words Feature	SVM	Late fusion of both acoustic and lexical features achieved four-class emotion recognition accuracy of 69.2%.

[25]	2015	CASIA, EmoDB, SAVEE, IEMOCAP, CHEAVD, AFEW	Spectral, Prosody and voice quality	SVM, Decision Tree, Random Forest, RBF, BayesNet	Prosodic and voice quality features are robust for emotion recognition on acted corpus, while spectral features are robust in induced and natural corpus.

TABLE III. A summary of dataset features and classifiers in 2016-2020

Literature	Year	Dataset	Features	Classifies	Conclusion
[74]	2016	RECOLA	Raw Audio Data	Convolutional Recurrent Network	The proposed model can outperform state-of-the-art model.
[75]	2017	EmoDB	Spectrogram	CNN	The trained model can predict emotions accurately and efficiently.
[76]	2017	EmoDB	Raw Audio Data	DNN	The trained model can achieve overall test accuracy of 96.97% on whole file classification.
[77]	2017	IEMOCAP	LLDs	RNN with local attention	The proposed model achieved better classification accuracy than traditional SVM-based models.
[78]	2017	FAU AIBO	Zero-Crossing Rate (ZCR), Root Mean Square (RMS), Energy, Pitch Frequency, HNR, MFCC	Ensemble Learning	The best unweighted average recall rate is 45.0% for the 5-class classifi cation task.
[16]	2018	EmoDB, IEMOCAP	Raw Audio Data, Log-mel Spectrogram	CNN LSTM	The 2-D CNN LSTM network achieves a recognition accuracy of 95.33% and 95.89% in speaker-dependent and speaker-independent experiments, respectively, which is better than traditional methods.
[79]	2018	RECOLA	Raw Audio Data	CNN and LSTM	Their model outperformed over the state-of-the-art methods in the RECOLA dataset.
[80]	2018	IEMOCAP	Spectrogram	Attention-based Bidirectional LSTM and FCN	The combination of BLSTM and FCN as well as the use of an attention mechanism can improve the performance on the IEMOCAP dataset.
[81]	2018	CASIA	spectrogram	CNN combined with Random Forest	CNN-RF model is superior to CNN model in recognition accuracy.
[82]	2018	EmoDB	Improved MFCC features, EEMFCC and	SVM	The highest average recognition rate of 85.37% for seven category emotions

			F0MFCC		and 100% for sadness are obtained.
[83]	018	IEMOCAP	Spectrogram	CNN	The proposed data preprocessing algorithm called DPARIP by augmenting the quantity of spectrograms is effective and more accurate compared to existing emotion recognition algorithms.
[84]	2018	CASIA, eNTERFACE, and GEMEP	Frame-level Features	LSTM with attention	the performance of the proposed approach outperformed the state-of-the-art algorithms reported to date.
[85]	2019	EmoDB, CASIA, IEMOCAP, CHEAVD	Raw Audio Data	CNN	the proposed algorithm is of great benefit to implement real-time speech emotion recognition.
[86]	2019	IEMOCAP	3 time-domain features, 5 spectral-domain features, 13 MFCCs, 13 Chroma	LSTM with attention	The results showed that the combination of noise elimination and attention model outperformed the use of either one of them.
[87]	2019	IEMOCAP, RAVDESS	Spectrogram	Deep Stride Convolutional Neural Network (DSCNN)	The recognition accuracy in RAVDESS is 79.5% and in IEMOCAP is 81.75%.
[88]	2020	CASIA, EmoDB, IEMOCAP	Raw Audio Data	Residual-CNN	The proposed algorithm provided significantly higher-accuracy predictions compared to existing speech emotion recognition algorithms in different language systems.
[89]	2020	IEMOCAP	Statistical Features and Log-mel Spectrogram	LSTM and CNN	Compared with the benchmark, the proposed model improved the ability of speech emotion modeling and effectively improved the accuracy of SER.

Deep learning is extremely popular worldwide in machine intelligence and it promoted speech emotion recognition greatly by establishing an end-to-end emotion recognition model, eliminating the hand-craft feature engineering in traditional machine learning algorithms. CNNs, RNNs and their variants, are such deep learning algorithms, and have been widely used in speech emotion recognition.

[76] developed an end-to-end deep learning framework to classify angry, sad, and neutral on EmoDB. They used a 320-dimensional feature vector to represent the fixed-length speech segment. Variable length speech utterances can be composed of different numbers of these fixed-length speech segments. However, only 33 emotional utterances were tested and emotion types were less. [75] input the spectrogram of the original speech signal directly into CNN, and relied on the powerful feature learning ability of CNN to extract robust and representative emotional features. In addition, they also introduced transfer learning and used the correlation between the two learning tasks to deal with insufficient training samples. The pre-trained AlexNet model is used for classification. But the results are not satisfactory. [83] used an

improved AlexNet for classification and based on retinal imaging, they performed data enhancement to generate multiple spectrograms for an emotional utterance. They conducted experiments on IEMOCAP and obtained an average recognition accuracy rate of 48.8%. [56] developed a CNN framework based on spectrograms and employed PCA whitening to carry out dimensionality reduction. They obtained a classification accuracy of 40% on IEMOCAP dataset, which proves to be more effective than SVM based on artificial features.

RNN is also widely used in emotion recognition due to its inherent contextual relevance between data. [96] established a BLSTM-RNN framework for classification of four primary emotions on IEMOCAP dataset with spectrogram features. They perform representation learning and transfer learning, which adapted a valence-activation trained RNN to a four-emotion classification task, to improve recognition performance. Besides, [6] verified that LSTM overperformed SVM in the prediction of dimensional emotions. [77] combined BLSTM with an attention mechanism to deal with the emotion-related segments of the speech.

It is natural to integrate multiple neural network models to investigate the performance of emotion recognition. CNN performs better at extracting spatial features, while RNN performs better at extracting temporal features. Therefore, CNN together with RNN can be a good candidate for emotion recognition. Zhao et al. [16] constructed a 1D CNN LSTM network and a 2D CNN LSTM network to learn local and global emotion-related features from speech and logmel spectrogram respectively. The combination of CNN and LSTM can inherit the strengths of both networks and overcome the shortcomings of them. Similarly, [74] combined CNNand LSTM to extract the context-aware emotional relevant features in order to automatically learn the best representation of the raw sequential audio data. The use of the proposed topology significantly outperforms the manual feature-based Support Vector Regression on RECOLA dataset. [80] constructed a novel network architecture, taking spectrogram as the model input, extracting spatial features with CNN, and modeling context dependence with BLSTM to capture temporal features, also applying attention mechanism to the salient parts of emotions, and then concatenate the extracted features and input them into the deep neural network to obtain higher prediction results. Compared to the state-of-art models, their features performed better, 65.2% for weighted accuracy and 68.0% for unweighted accuracy.

Deep neural networks such as CNN, LSTM, and their combinations can effectively extract deep features and have achieved competitive performance for automatic speech emotion recognition compared with hand-crafted and tailored features. Many of these experiments using deep learning models are end-to-end models that bypass the manual feature extraction that relies on expert knowledge in traditional machine learning. Fig. 5 is an prototype of such an end-to-end automatic speech emotion recognition architecture.

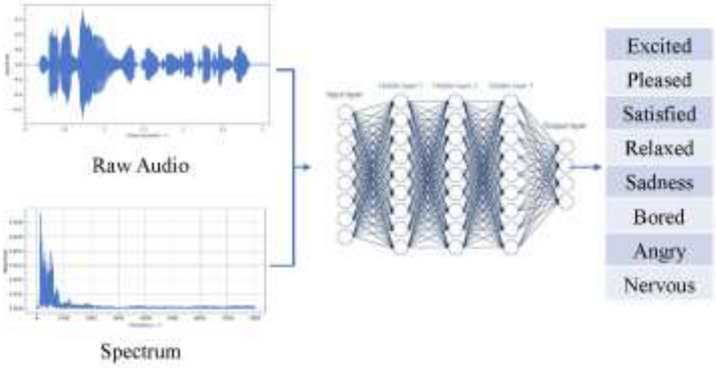

Fig 5: End-to-end automatic speech emotion recognition

Coupling traditional machine learning with deep neural networks is promising and lucrative. [72] investigated such a coupled DNN-HMM model. And they demonstrated that the DNN could extend the labeling ability of GMM-HMM with a best recognition accuracy of 53.89%. [81] combined convolution neural network with Random Forest (RF) to extract appropriate emotion features on the CASIA dataset. CNN acts as a feature extractor to extract spectrogram features, which were input to the random forest classifier. The results showed that their CNN-RF model is superior to CNN in terms of the classification accuracy.

We have compiled and summarized some important information about datasets, features, model selection, etc. More details between 2000 and 2015 are shown in Table II. Details between 2016 and 2020 are shown in Table III. The information may be useful for future speech emotion recognition. A basic conclusion can be drawn that from 2000 to 2015, traditional machine learning methods occupy a dominant position in the selection of models, while between 2016 and 2020, researchers are more inclined to use deep learning methods, such as CNN and RNN. It is difficult to figure out which classifier and model have the best performance. Most of the above literatures have proved that methods based on deep learning are better than traditional machine learning methods. But there are also some literatures that contradict this view. Due to the inconsistency of evaluation standards and the dependence of speech emotion recognition on specific dataset, these methods are incomparable. Undoubtedly, these works have promoted the continuous forward development of speech emotion recognition research to a certain extent.

IV. FACTORS AFFECTING SPEECH EMOTION RECOGNITION

4.1 Gender Discrepancy

Many researches show that gender has a great impact on the accuracy of speech emotion recognition. According to [97], women speak faster than men when they are angry, which may bias the experimental results. In [17], Liu et al. showed that the proposed method improved by 9.0% on the female model and 7.4% on the male model compared to the benchmark. The author of [55] conducted a gender-dependent experiment, and the recognition accuracy of the male model was higher than that of the female model, since women were more inclined to express emotions in multiple ways such as gestures. The same conclusion was also drawn by You et al. [57].

The authors of [92] have done gender-dependent emotion recognition and gender-independent emotion recognition respectively, and experiments showed that the gender-dependent had higher recognition accuracy. Regarding gender as an affecting factor, in actual emotion recognition applications, gender-dependent emotion recognition can obtain higher accuracy. More representative and significant emotion related acoustic features should be further explored instead of gender related.

4.2 Cultural Backgrounds

People with different cultural backgrounds usually have very different emotional expressions. Some people tend to express happiness with high loudness, others in a quieter way. The current research on multilingual speech emotion recognition is relatively rare. Therefore, constructing a speech emotion dataset containing speakers with many different cultural backgrounds is a good solution to this problem. Experiments on existing data sets to eliminate the influence of cultural background are essential to achieve high-performance speech emotion recognition. The extraction of features related to culture background is also worth exploring.

4.3. Noise Interference

The environment noise seriously interferes with the speech signals, which greatly reduces the recognition accuracy. Three major ways to reduce noise interference: robust feature extraction, model adaptation, and speech enhancement [91]. Spectrum subtraction is an effective speech enhancement algorithm. It adds noise to the training process of the denoising autoencoder to obtain a more efficient feature representation, which can be used as the input of the classifier to improve system performance and reduce noise interference.

4.4. Model Selection

The choice of classifier is critical to classification accuracy. It is generally believed that SVM is suitable for processing high-dimensional voice data, which does not depend on the data amount, and has a strong generalization ability. The choice of the kernel function also has a great impact on the recognition results. For deep neural networks, hyperparameters setting is a process of constant experimentation, such as the learning rate, the number of neurons in each layer, etc. Both over-fitting and under-fitting will affect the recognition accuracy. In [72], a model based on DNN-HMM is proposed to classify emotions. They studied the influence of the number of hidden layers of DNN and found that when the number of hidden layers of DNN exceeds 6, the performance of the model will deteriorate. Therefore, the importance of a stable sentiment classification model and appropriate hyperparameters is self-evident.

V. FUTURE DIRECTIONS

At present, speech emotion recognition has achieved fruitful results. A major problem is that most recognition algorithms are based on a single emotion speech dataset. In other words, the recognition performance is greatly reduced when samples derived from other emotion datasets are tested. The generalization ability of the model needs to be further improved. In essence, the extracted features are not sufficiently representative. And the model may have learned some features that are unrelated to emotions. A suggestion is to train the model on multiple datasets. At present, research in this area is relatively scarce. Building a large, multilingual speech emotion dataset may also be a feasible direction.

Most speech emotion recognition algorithms are supervised. The performance cannot be guaranteed, relying on hand-craft and tailored features. Therefore, the lack of labeled training data would be a concentrated problem. Effective data enhancement algorithms can be used to expand the amount of available data. For example, based on the principle of retinal imaging, each emotional utterance can generate multiple spectrograms, effectively expanding the training set [83].

In addition, some researchers try to use semi-supervised algorithms. In [17], in order to effectively use unlabeled data, the author uses a semi-supervised collaborative training algorithm. Other scholars try to use transfer learning to solve the problem of insufficient labeled data [98].

Due to the lack of a unified standard, different people may have different understandings of the same emotion utterance, which leads to the labeled data being very subjective. Therefore, the emotion theory must be further explored. The classification of emotions in many current studies only involves very simple emotions. [96] only discussed the classification of four basic emotions: neutral, anger, sadness and happiness. Few studies have focused on complex emotions such as disgust and anxiety.

The choice of acoustic features is crucial in speech emotion recognition. Feature selection and extraction for representative and compact feature representations are still a major direction. Another major problem of deep learning is the poor interpretability of the model. The breakthroughs of the above two research interests may be a milestone in the field of speech emotion recognition.

In general, researchers are more inclined to the fusion of multiple acoustic features [82], the decision fusion of multiple classifiers[69], the combination of deep learning and traditional machine learning[81,72]. However, most of the relevant research stays in the laboratory experiment stage, and the real-time performance and practical application are not ideal enough, which requires continuous exploration.

Article History: Received: 05 October 2021, Revised: 02 November 2021, Accepted: 25 November 2021, Publication: 31 December 2021

V. CONCLUSION

Speech emotion recognition is a popular topic with various applications from human-computer interfaces to affective computing. In this article, we have systematically summarized commonly used emotion datasets, various feature extraction and dimensionality reduction algorithms, and some mainstream model frameworks for the previous emotion recognition research. And we also summarized and compared traditional machine learning and deep learning algorithms used for emotion recognition, and gave suggestions on existing problems and future research directions.

ACKNOWLEDGEMENTS

This work was supported by the National Nature Science Foundation of China under Grant U1633121, and University of Science and Technology Course Fund KC2019SZ11 and 54896055.

REFERENCES

[1] Park JS, Kim JH, and Oh YH (2009) Feature vector classification based speech emotion recognition for service robot. IEEE Transactions on Consumer Electronics 55(3):1590–1596

[2] Li Y, Ishi CT, Ward N, Inoue K, Nakamura S, Takanashi ., Kawahara T (2017) Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot. in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE 1356–1359

[3] Jansen MP (2019) Communicative signals and social contextual factors in multimodal affect recognition. in 2019 International Conference on Multimodal Interaction 468–472

[4] Ekman P (1992) An argument for basic emotions. Cognition & emotion 6(3-4):169–200

[5] Gunes H, Pantic M. (2010) Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions (IJSE) 1(1):68–99

[6] Chen S, Jin Q, Zhao J, and Wang S (2017) Multimodal multi-task learning for dimensional and continuous emotion recognition. in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 19– 26

[7] Faria DR, Vieira M, Faria FC (2017) Towards the development of affective facial expression recognition for human-robot interaction. in Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments 300–304

[8] Zhang L, Hossain A, Jiang M (2014) Intelligent facial action and emotion recognition for humanoid robots. in 2014 International Joint Conference on Neural Networks (IJCNN) IEEE 739–746

[9] Lopez-Rincon A (2019) Emotion recognition using facial expressions in children using the nao robot. in 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP) IEEE 146– 153

[10] Juszkiewicz Ł (2014) Improving speech emotion recognition system for a social robot with speaker recognition. in 2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR), IEEE 921–925

[11] Anjum M (2019) Emotion recognition from speech for an interactive robot agent. in 2019 IEEE/SICE International Symposium on System Integration (SII) IEEE 363–368

[12] Le BV, Lee S (2014) Adaptive hierarchical emotion recognition from speech signal for human-robot communication. in 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing IEEE 807–810

[13] Keshari T, Palaniswamy S (2019) Emotion recognition using feature-level fusion of facial expressions and body gestures. in 2019 International Conference on Communication and Electronics Systems (ICCES) IEEE 184–1189

[14] Nguyen D, Nguyen DT, Zeng R, Nguyen TT, Tran SN, Nguyen T, Sridharan S, Fookes C (2021) Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. IEEE Transactions on Multimedia

[15] Chen G, Zhang X, Sun Y, Zhang J (2020) Emotion feature analysis and recognition based on reconstructed eeg sources. IEEE Access 8:11907–11916

[16] Zhao J, Mao X, and Chen L (2019) Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomedical Signal Processing and Control 47:312–323

[17] Liu J, Chen C, Bu J, You M, and Tao J. (2007) Speech emotion recognition using an enhanced co-training algorithm. in 2007 IEEE International Conference on Multimedia and Expo IEEE 999–1002

[18] Kim DH (2013) Fuzzy rule based voice emotion control for user demand speech generation of emotion robot. in 2013 International Conference on Computer Applications Technology (ICCAT) IEEE 1–4

[19] Bieber G, Antony N, and Haescher M (2018) Touchless heart rate recognition by robots to support natural human-robot communication. in Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference 415–420

[20] Shen P, Changjun Z, Chen X (2011) Automatic speech emotion recognition using support vector machine. in Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology IEEE 2:621–625

[21] Le D, Provost EM (2013) Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding IEEE 216–221

[22] Lin J, Pan S, Lee CS, Oviatt S (2019) An explainable deep fusion network for affect recognition using physiological signals. in Proceedings of the 28th ACM International Conference on Information and Knowledge Management 2069–2072

[23] Yang L, Xie K, Wen C, He JB (2021) Speech emotion analysis of netizens based on bidirectional lstm and pgcdbn. IEEE Access 9:59860–59872

[24] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. in Ninth European Conference on Speech Communication and Technology

[25] Li Y, Chao L, Liu Y, Bao W, Tao J (2015) From simulated speech to natural speech, what are the robust features for emotion recognition? in 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) IEEE 368–373

[26] Zhu C, Ahmad W (2019) Emotion recognition from speech to improve human-robot interaction. in 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) 370–375

[27] Zhang JTFLM, Jia H (2008) Design of speech corpus for mandarin text to speech. in The Blizzard Challenge 2008 workshop

[28] Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a danish emotional speech database. in Fifth European conference on speech communication and technology

[29] Ringeval F, Sonderegger A, Sauer J, Lalanne D (2013) Introducing the recola multimodal corpus of remote collaborative and affective interactions. in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) 1–8

[30] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42(4):335

[31] Steidl S (2009) Automatic classification of emotion related user states in spontaneous children's speech. Logos-Verlag 250

[32] Rabiei M, Gasparetto A (2016) System and method for recognizing human emotion state based on analysis of speech and facial feature extraction; applications to human-robot interaction. in 2016 4th International Conference on Robotics and Mechatronics (ICROM) 266– 271

[33] Rabiner LR (1978) Digital processing of speech signal. Digital Processing of Speech Signal

[34] Hernando J, Nadeu C (1997) Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. IEEE Transactions on Speech and Audio Processing 5(1):80–84

[35] Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing 28(4):357–366

[36] Atal BS (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America 55(6):1304– 1312

[37] Schmitt M, Ringeval F, Schuller BW (2016) At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. in Interspeech 495–499

[38] Huang J, Li Y, Tao J, Lian Z, Niu M, Yi J (2018) Speech emotion recognition using semi-supervised learning with ladder networks. in 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia) IEEE 1–5

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

[40] Zhou Y, Sun Y, Zhang J, Yan Y (2009) Speech emotion recognition using both spectral and prosodic features. in 2009 International Conference on Information Engineering and Computer Science IEEE 1–4

[41] Stevens SS, Volkmann J, Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America 8(3):185–190

[42] Basu S, Chakraborty J, Aftabuddin M (2017) Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. in 2017 2nd International Conference on Communication and Electronics Systems (ICCES) 333-336

[43] Likitha MS, Gupta SR, Hasitha RK, Raju AU (2017) Speech based human emotion recognition using mfcc. in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) 2257–2260

[44] Umamaheswari J, Akila A (2019) An enhanced human speech emotion recognition using hybrid of prnn and knn. in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) 177–183

[45] Lanjewar RB, Chaudhari D (2013) Speech emotion recognition: a review. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2(4):68–71

[46] Luengo I, Navas E, Herna´ez I (2010) Feature analysis and evaluation for automatic emotion identification in speech. IEEE Transactions on Multimedia 12(6):490–501

[47] Lugger M, Yang B (2007) The relevance of voice quality features in speaker independent emotion recognition. in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 4:IV–17

[48] Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE transactions on multimedia 16(8):2203–2213

[49] Jin Q, Li C, Chen S, Wu H (2015) Speech emotion recognition with acoustic and lexical features. in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) IEEE 4749–4753

[50] Eyben F, Wo¨llmer M, Schuller B (2010) Open smile: the munich versatile and fast open-source audio feature extractor. in Proceedings of the 18th ACM international conference on Multimedia 1459–1462

[51] Chao L, Tao J, Yang M, Li Y (2014) Improving generation performance of speech emotion recognition by denoising autoencoders. in the 9th International Symposium on Chinese Spoken Language Processing IEEE 341–344

[52] Xu X, Li Y, Xu X, Wen Z, Che H, Liu S, Tao J (2014) Survey on discriminative feature selection for speech emotion recognition. in the 9th International Symposium on Chinese Spoken Language Processing IEEE 345–349

[53] Pudil P, Ferri FJ, Novovicova J, Kittler J (1994) Floating search methods for feature selection with nonmonotonic criterion functions. in Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5) IEEE 2:279–283

[54] Ververidis D, Kotropoulos C, Pitas I (2004) Automatic emotional speech classification. in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing 1:I–593

[55] You M, Chen C, Bu J, Liu J, Tao J (2006) A hierarchical framework for speech emotion recognition. in 2006 IEEE international symposium on industrial electronics 1:515–519

[56] Zheng W, Yu J, Zou Y (2015) An experimental study of speech emotion recognition based on deep convolutional neural networks. in 2015 international conference on affective computing and intelligent interaction (ACII) IEEE 827–831

[57] You M, Chen C, Bu J, Liu J, Tao J (2006) Emotional speech analysis on nonlinear manifold. in 18th International Conference on Pattern Recognition (ICPR'06) 3:91–94

[58] Tenenbaum J De Silva BV, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. science 290(5500):2319–2323

[59] Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500):2323–2326

[60] Xia R, Liu Y (2013) Using denoising autoencoder for emotion recognition. in Interspeech 2886–2889

[61] Nwe TL, Foo SW, and De Silva LC (2003) Speech emotion recognition using hidden markov models. Speech communication 41(4):603–623

[62] Vondra M, V´ıch R (2009) Evaluation of speech emotion classification based on gmm and data fusion. in Cross-modal analysis of speech, gestures, gaze and facial expressions, Springer 98–105

[63] Tsai CC, Chen YZ, Liao CW (2009) Interactive emotion recognition using support vector machine for human-robot interaction. in 2009 IEEE International Conference on Systems, Man and Cybernetics 407–412

[64] LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

[65] Kombrink S, Mikolov T, Karafia´t M, Burget L (2011) Recurrent neural network based language modeling in meeting recognition. in Twelfth annual conference of the international speech communication association

[66] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780

[67] Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

[68] Lin YL, Wei G (2005) Speech emotion recognition based on hmm and svm. in 2005 international conference on machine learning and cybernetics, IEEE 8:4898–4901

[69] Bozkurt E, Erzin E, Erdem CE, Erdem AT (2010) Use of line spectral frequencies for emotion recognition from speech. in 2010 20th International Conference on Pattern Recognition, IEEE 3708–3711

[70] Chavhan Y, Dhore M, Yesaware P (2010) Speech emotion recognition using support vector machine. International Journal of Computer Applications 1(20):6–9

[71] Pan Y, Shen P, Shen L (2012) Speech emotion recognition using support vector machine. International Journal of Smart Home 6(2):101–108

[72] Li L, Zhao Y, Jiang D, Zhang Y, Wang F, Gonzalez I, Valentin E, Sahli H (2013) Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. in 2013 Humaine association conference on affective computing and intelligent interaction, IEEE 312– 317

[73] Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using fourier parameters. IEEE Transactions on affective computing 6(1):69–75

[74] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S (2016) Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 5200–5204

[75] Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. in 2017 international conference on platform technology and service (PlatCon), IEEE 1–5

[76] Hara´r P, Burget R, Dutta MK (2017) Speech emotion recognition with deep learning. in 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE 137–140

[77] Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE 2227–2231

[78] Shih PY, Chen CP, Wu CH (2017) Speech emotion recognition with ensemble learning methods. in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE 2756–2760

[79] Tzirakis P, Zhang J, Schuller BW (2018) End-to-end speech emotion recognition using deep neural networks. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE 5089–5093

[80] Zhao Z, Zhao Y, Bao Z, Wang H, Zhang Z, Li C (2018) Deep spectrum feature representations for speech emotion recognition. in Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data 27–33

[81] Zheng L, Li Q, Ban H, Liu S (2018) Speech emotion recognition based on convolution neural network combined with random forest. in 2018 Chinese Control and Decision Conference (CCDC), IEEE 4143– 4147

[82] Wang Y, Hu W (2018) Speech emotion recognition based on improved mfcc. in Proceedings of the 2nd international conference on computer science and application engineering 1–7

[83] Niu Y, Zou D, Niu Y, He Z, Tan H (2018) Improvement on speech emotion recognition based on deep convolutional neural networks. in Proceedings of the 2018 International Conference on Computing and Artificial Intelligence 13–18

[84] Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B (2019) Speech emotion classification using attention-based lstm. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(11):1675–1685

[85] Gao M, Dong J, Zhou D, Zhang Q, Yang D (2010) End-to-end speech emotion recognition based on one-dimensional convolutional neural network. in Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence 78–82

[86] Atmaja BT, Akagi M (2019) Speech emotion recognition based on speech segment using lstm with attention model. in 2019 IEEE International Conference on Signals and Systems (ICSigSys), IEEE 40–44

[87] Latif S, Rana R, Khalifa S, Jurdak R, Epps J (2019) Direct modelling of speech emotion from raw speech. arXiv preprint arXiv:1904.03833

[88] Sun TW (2020) End-to-end speech emotion recognition with gender information. IEEE Access 8:152423–152438

[89] Huilian L, Weiping H, Yan W (2020) Speech emotion recognition based on blstm and cnn feature fusion. in Proceedings of the 2020 4th International Conference on Digital Signal Processing 169–172

[90] Zhiyan H, Jian W (2013) Speech emotion recognition based on wavelet transform and improved hmm. in 2013 25th Chinese Control and Decision Conference (CCDC), IEEE 3156–3159

[91] Chenchah F, Lachiri Z (2016) .Speech emotion recognition in noisy environment," in 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE 788–792

[92] Hu H, Xu MX, Wu W (2007) Gmm supervector based svm with spectral features for speech emotion recognition. in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, IEEE 4:IV–413

[93] Lasiman JJ, Lestari DP (2018) Speech emotion recognition for indonesian language using long short-term memory. in 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE 40–43

[94] Lee CC, Mower E, Busso C, Lee S, Narayanan S (2011) Emotion recognition using a hierarchical binary decision tree approach. Speech Communication 53(9-10):1162–1171

[95] Kim Y, Provost EM (2013) Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing 3677–3681

[96] Ghosh S, Laksana E, Morency LP, Scherer S (2016) Representation learning for speech emotion recognition. in Interspeech 3603–3607

[97] Heuft B, Portele T, Rauth M (1996) Emotions in time domain synthesis. in Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, IEEE 3:1974–1977

[98] Ottl S, Amiriparian S, Gerczuk M, Karas V, and Schuller B (2020) Group-level speech emotion recognition utilizing deep spectrum features. in Proceedings of the 2020 International Conference on Multimodal Interaction 821–826.