2021 International Wireless Communications and Mobile Computing (IWCMC) | 978-1-7281-8616-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IWCMC51323.2021.9498858

A survey on the development of intelligent robots in speech emotion recognition

Qingnan Gao, Huansheng Ning, Bing Du*
School of Computer and Communication Engineering, University of Science and Technology Beijing
s20200645@xs.ustb.edu.cn, ninghuansheng@ustb.edu.cn, dubing@ustb.edu.cn

***Abstract*—Speech emotion recognition, an important branch of affective computing, has attracted much attention recently, which is of great significance for the realization of natural and harmonious human-robot interaction. Many researches have been carried out and remarkable results have been achieved in this field. At the same time, there are still many problems to be tackled urgently. Therefore, it is necessary to summarize the previous works and find out the problems in speech emotion recognition, so as to provide guidance for further research. This paper systematically summarizes the general process of speech emotion recognition, including speech emotional databases and various leading emotion classification models. By listing some outstanding works of speech emotion recognition in recent 20 years, this paper compares and analyzes the highlights and shortcomings of these works. It can be seen that people show more interest in deep learning in which features are mostly extracted automatically than the traditional machine learning methods. Finally, the main problems in the field of speech emotion recognition and the direction of further exploration are summarized in order to promote speech emotion recognition to a new stage.**

***Index Terms*—affective computing, speech emotion recognition, emotion classification models**

I. Introduction

In recent years, artificial intelligence has become more and more prevalent in the field of computer science. The rapid renovation of computer hardware and the proposal of novel neutral network models are main reasons. As one of the application fields of artificial intelligence, intelligent robots show the characteristics of the new era. The most popular and widely used intelligent robots are service robots which are better at doing dangerous and sophisticated tasks than humans because of their extremely fast calculation speed and high precision. Therefore, service robots applied in the areas such as smart medical care, health care and home service achieve extraordinary outcomes [1]. However, intelligent robots have little emotion, which makes the interaction between humans and robots not so smooth and natural [2]. In view of this, many people devote themselves to the research of emotional robots which have the capacity to understand human emotion and give appropriate response.

The conception of emotional robots also called affective computing and artificial psychology was first proposed by Professor Picard of MIT Media Laboratory in the 1990s [3]. Furthermore, using varieties of emotional expressions (e.g. speech signals, facial expressions, body gestures) to automatically recognize emotion is one of research branches derived from affective computing, which has attracted extensive attentions in recent years [4]. However, human emotion is abstract and intangible, so that one may fail to distinguish and understand other people's emotion conveyed by his/her behaviors, let alone intelligent robots. In order to thoroughly analyze human emotion, two emotional theories called categorial emotion and dimensional emotion are put forward. The categorial emotional theory regards emotion as various isolated and discrete states and contends that human emotion can be divided into seven categories, namely happiness, anger, disgust, fear, sadness, surprise and neutral [5]. Instead of discretizing emotion, the dimensional emotional theory confirms that emotion is continuous and can be placed in valence-arousal space. Valence is measurement of the positive and negative of emotion, while arousal represents the level of the emotional activation [6]. Based on the above traits, many complex and subtle expressions can be modeled by the dimensional emotion theory [7]. Fig. 1 shows the dimensional emotion model described by valence and arousal. Both of the theories are generally accepted by people and applied to the research of the emotion recognition of human-robot interaction (HRI).

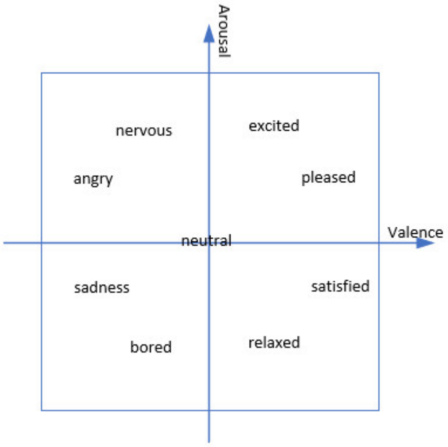

Fig. 1. The model of dimensional emotion

This paper mainly focuses on the speech emotion recognition. Speech emotion recognition usually refers to infer the speaker's emotional state from speech. Speech is one of the most direct channels used for communication in which people can perceive each other's emotional changes through speech

This work is supported by Civil Aviation Joint Funds of the National Natural Science Foundation of China (GrantNo.U1633121)

signals [8]. Endowing intelligent robots with the ability to automatically recognize human emotional state from speech signals is an essential component of the affective computing [9]. However, there are not consistent views about which dimensional reduction methods should be utilized to effectively produce compact and representative acoustic features in high-dimensional speech data with noisy [10]. How to optimize speech emotion recognition architecture called end-to-end learning which utilizes the technology of deep learning and improve the models explainability should also be further addressed [11]. Considering the above serious situation, it is imperative to summarize previous research outcomes and promote speech emotion recognition to a new stage.

The remainder of this paper is arranged as follows. Section II focuses on the leading approaches of speech emotion recognition. Section III mainly investigates and analyzes the methods and related work done by the researchers in the speech emotion recognition. Next, the existing problems and the direction of exploration are shown in Section IV. Section V concludes the paper.

II. The approaches of speech emotion recognition

A. Support Vector Machine (SVM)

SVM is a canonical and discriminant model, which is extensively used in the field of speech emotion recognition. Although SVM is not capable to fit the data distribution like GMM, SVM has some other remarkable advantages, such as perfect adaption for small sample learning, insusceptible to data dimension, high accuracy of classification, solid theoretical foundation etc.

Assuming that the training set contains n samples, namely $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $y \in \{-1, 1\}$, linear SVM attempts to find a hyperplane, which is competent to optimally separate the two types of samples [12]. The hyperplane equation can be described as:

$$w^T x + b = 0 \tag{1}$$

The objective function of optimization is:

$$\max_{w,b} \frac{2}{||w||} \tag{2}$$

The constraint is:

$$y_i(w^T x_i + b) \geq 1 \tag{3}$$

The two parameters w and b in the hyperplane equation can be easily obtained through efficient optimization algorithms. Ultimately, the hyperplane equation can be determined. It is noted that this is only the case where samples are linearly separable. Considering the complexity of speech features, it is difficult for linear SVM to distinguish different emotions effectively. When it comes to the case where samples are linearly inseparable, kernel function addresses this problem by mapping feature vectors from original space to high dimensional space to search for a optimal separating hyperplane. Kernel function is the inner product of two vectors mapped to high dimensional space. Simultaneously, the introduction of kernel function also reduces the computational cost. The commonly used kernel functions include linear kernel, Gaussian kernel and so on. Equation (4) is linear kernel, while (5) is Gaussian kernel.

$$K(x_i, x_j) = x_i^T x_j \tag{4}$$

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{5}$$

Considering the variety of emotions, it is necessary to employ multi-SVM. There are two generic approaches about multi-SVM consisting of one-to-one and one-to-rest. Assuming that there are k kinds of emotions, the former establishes $k(k-1)/2$ hyperplane equations, while the latter establishes k hyperplane equations.

B. Recurrent Neural Network (RNN) and its variants

As an integral part of deep neural network, RNN has been received more attention by many researchers for machine translation and speech recognition, because RNN possesses the edge over processing sequence data and achieves state-of-the-art performance. Instead of treating the input data as isolated and irrelevant, RNN attempts to extract the contextual information of the sequence, so it can capture more temporal information, which is critically important for audio signal and other sequence data. It is worth mentioning that RNN is also a weight sharing mode, which is reflected in using the same weight matrix at different times. Fig. 2 shows the structure of original RNN. Equation (6) is mathematical formula of hidden states.

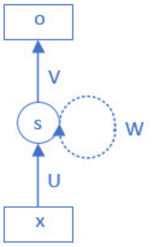

Fig. 2. The structure of original RNN

$$s_t = g(U \cdot x_t + W \cdot s_{t-1}) \tag{6}$$

The output of the current time can be computed by:

$$o_t = f(V \cdot s_t) \tag{7}$$

$U, V,$ and W are weight matrices and g, f are nonlinear activation function such as $Sigmoid$ and $Relu$. x_t denotes the input of current time t. s_t represents the state of hidden layers. o_t is the output of the current time t. It can be seen that the key point of RNN is to utilize the input of the current time x_t integrated with the hidden state of the previous time s_{t-1} to collectively predict the output of the current time o_t [13]. Moreover, unfolding on the time line, RNN is competent to effectively address the problem of uncertain length of sequence data , which is complex to resolve for traditional machine learning.

One of the problems of RNN is that the gradient vanish, leading to some parameters can not be updated effectively in the back propagation through time (BPTT) algorithm. Therefore, a variant of RNN, called long short-term memory (LSTM) [14], is proposed. LSTM ameliorates the structure of RNN by introducing the concept of cell state as well as gate mechanism including forgetting gate, input gate and output gate. Although LSTM can overcome the long-term dependence problem of RNN, it is very disadvantageous for the design of real-time system to increase the amount of computation.

LSTM achieves higher performance in conjunction with attention mechanism [15], when confronting with extremely long sequence data, thus drawing a great attraction. Attention mechanism mimics the process that human often focus on local area of the target to learn better about it. In the research of speech emotion recognition, since different segments of the emotional utterance are unlike in emotional saturation, more attention will be paid to the parts with high emotional saturation. Equation (8) shows the calculation process of the weight in each time step t.

$$\alpha_t = \frac{exp(u^T y_t)}{\sum\limits_{i=1}^{T} exp(u^T y_i)} \tag{8}$$

In the (8), u represents the attention parameter vector. and y_i corresponds to the output in time step i. Inner product operation between u^T and y_i is to measure the similarity between them. Then weight denoted as α_t can be obtained with a softmax function. Finally, the context vector c can be calculated by (9) in a form of weighted sum.

$$c = \sum_{t=1}^{T} \alpha_t y_t \tag{9}$$

III. Discussion about which way has the optimal performance, traditional machine learning or deep learning

The use of manual features as input of traditional machine learning classifiers has been in the field of speech emotion recognition for a long time. Typical methods consist of HMM, GMM, SVM, decision trees, KNN, naive Bayes, etc.

In [16], Zhou et al. designed a GMM supervector based SVM architecture with spectral features as the input of the model. Similar to this. Hu et al. [17] also used the SVM architecture based on Gaussian supervector, and also used the GMM model for comparative analysis. The SVM based on Gaussian supervectors proved to be more effective than GMM in gender-dependent systems. The proposed method has a higher classification accuracy of 5.7%, in the male sample and 6.3% in the female sample than baseline. In [18], You et al. trained a linear support vector machine classifier. The input of the model is a six-dimensional abstract feature representation after dimensionality reduction with an enhanced Lipschitz embedding. In the speaker-independent experiment, a relative improvement of 9%-26% was achieved. Similarly, they carried out a speaker-dependent experiment, which resulted in a relative accuracy improvement of 5%-20%. Wang et al. [19] used the SVM classifier on the EMODB dataset [20], and obtained the highest classification accuracy of 85.37% using two improved MFCC features. They noticed that the recognition rate of happiness was only 50%. Further, analyzing the confusion matrix can find that many samples of happiness are incorrectly classified as anger. They argued that the two emotions of happiness and anger were too similar and both of them belonged to the category of high arousal. Combining multiple features can effectively improve the accuracy of sentiment classification. Shen et al. [21] used a variety of acoustic features including prosody and spectral features as the input of SVM, and concluded that combining multiple features was more advantageous than a single feature, which was consistent with the viewpoint of [19]. In addition, SVM is often used as the baseline to judge other models [22] [23]

In addition to SVM, other machine learning methods are also fully used in speech emotion recognition and have achieved extraordinary performance. Schmitt et al. [24] proposed a support vector regression based on bag-of-audio-words (BoAW) representations to deal with the problem of dimensional emotion prediction. On the RECOLA dataset [25], an important conclusion that the framework they proposed even dropped some emotion recognition models based on CNN and RNN so far was drawn. In [26], a robust decision tree system was constructed for speech emotion recognition. Their model has an absolute performance improvement of 7.44% over the SVM benchmark model on the IEMOCAP dataset [27]. KNN is also used as a classifier for speech emotion recognition in [28]. Bayesian classifiers are supported by profound mathematical theory and are not sensitive to missing data, which has attracted the interest of many researchers. A Bayesian classifier was adopted by Wang et al. [29]. Compared with the baseline, their recognition accuracy was improved by 6.01%.

In [30], in response to uneven distribution of emotional utterances in FAU AIBO [31], the author decided to apply the ensemble learning method. The unweighted average recall rate (UA) was used as the performance evaluation, and the best recognition accuracy was 45%. Rather than single-stage emotion classification, [32] pioneered the idea of cascading strategy. Based on a general consensus, prosodic features are good at distinguishing emotions of different arousal levels, and adopted a two-stage classification method. In the first stage, they divided emotions into two categories according to different activation levels. In the second stage, for each category in the first stage, they further subdivided the true categories of emotions. In this way, they obtained an average recognition rate of 74.5%. The main reasons behind the superior classification performance of the cascade strategy may be the following two points. First of all, in the first stage, the prosody feature is very distinguishable for the activation of emotions, which leads to a very high recognition accuracy in the first stage, reaching an average accuracy of 95.5%. Furthermore, in the second stage, for each category in the first stage, the emotional category interval needs to be further subdivided and the emotional category is reduced, which helps to improve the classification accuracy.

From the above speech emotion recognition based on machine learning, the following development trends can be summarized. Researchers tend to use a variety of acoustic features, while applying various feature selection and dimensionality reduction methods to extract representative and discriminative feature representations. In the selection of models, some choose a single robust model, and some adopt a classifier-level decision fusion strategy. In other words, in traditional machine learning, the selection of acoustic features and the construction of models are crucial steps for an efficient and robust speech emotion recognition system. Fig. 3 shows the process of speech emotion recognition based on traditional machine learning methods.

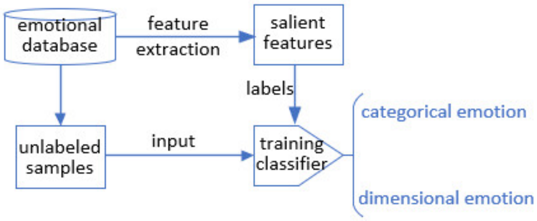

Fig. 3. Speech emotion recognition using traditional machine learning methods

In recent years, as an emerging technology, deep learning has been widely used in speech emotion recognition and is highly sought after by researchers. Convolutional neural networks, recurrent neural networks and their variants, as the main representative works of deep learning, have attracted extensive attention in the field of speech emotion recognition. In [33], the authors employed an end-to-end deep learning framework to classify three emotions on EMODB, including angry, sad, and neutral. Because of the variability of speech length, they divided emotional speech into many fixed-length segments, and each segment can be represented by a 320-dimensional feature vector as the input of the model. It is encouraging that they pioneered the adoption of an end-to-end architecture, which greatly reduced human labor, supported by powerful computing power and massive data. At the same time, the limitation is that the types of emotions classified are not enough. In the testing phase, the average probability of the prediction results of each segment corresponding to an emotional utterance needs to be used as a measure of the final result. There are only 33 emotional utterances to be tested, thus leading the experimental results may not be too convincing. In [34], spectrograms derived from the original speech signal are directly used as the input of CNN, relying on the powerful feature learning ability of CNN to extract robust and representative emotional features. Another highlight is that they introduced the theory of transfer learning into speech emotion recognition, which used the correlation between the two types of learning tasks to solve the problem of insufficient training samples to a certain extent. The pre-trained AlexNet model is used for classification. The experimental results of transfer learning are not satisfactory. Niu et al. [35] did a similar work to that in [34]. They used a modified AlexNet to make a classification model, and also used a spectrogram as input. They experimented on IEMOCAP and obtained an average accuracy of 48.8%. Another highlight is that they used data enhancement methods to generate multiple spectrograms for an emotional utterance through an algorithm based on the imaging principle of the retina. Zheng et al. [22] also adopted a spectrogram-based CNN framework. The improvement is that PCA whitening is applied to the spectrogram to achieve dimensionality reduction. Comparative experimental analysis showed that the classification accuracy of 40% was obtained on the IEMOCAP dataset and proved to be more effective than SVM based on manual features.

In addition to these CNN-based speech emotion recognition models, RNN is also widely used due to the advantages of considering the contextual relevance between data and being able to extract temporal information from emotional utterances. Ghosh et al. [36] used the BLSTM framework for four-class emotion recognition on the IEMOCAP dataset. In view of the fact that many emotional utterances on this dataset are marked by the valence and arousal, they decided to adopt the idea of transfer learning and used these emotional utterances to pre-train the network. This has two advantages. On the one hand, a large number of emotional utterances that do not belong to category emotion labels are fully used. On the other hand, the relationship between dimensional emotion and category emotion can be further explored. In the analysis of experimental results, a confusing problem is that the happy class is always incorrectly classified as the anger class, which is consistent with [19]. Another conclusion is that transfer learning has little effect on the improvement of recognition accuracy. In [7], experiments show that compared with traditional support vector regression, LSTM with powerful sequence modeling capabilities has better generalization capabilities in dimensional emotion prediction. The work of [37] uses BLSTM with an attention mechanism to make emotion-related parts get more attention.

Similar to the fusion of multiple classifiers in machine learning, combining multiple neural network models has received a lot of attention. In particular, CNN is good at extracting spatial features, and RNN can learn time domain information from sequence streams. The combination of the two can significantly improve the accuracy of emotion recognition. Zhao et al. [8] designed two network architectures to integrate CNN and LSTM, one 1D CNN LSTM network is used to learn from raw audio data, and one 2D CNN LSTM network is used to learn from manual features. Similarly, Trigeorgis et al. [38] proposed a convolutional recurrent neural network based on CNN and LSTM, taking raw audio data as input, and establishing an end-to-end dimensional emotion recognition system. Research shows that the proposed method surpasses the manual feature based SVR on the RECOLA dataset. Another pioneering work is that they directly use the performance evaluation standard concordance correlation coefficient (CCC) as the loss function in the training phase instead of mean square error (MSE). The authors of [39] constructed a novel network architecture using spectrograms as model input. The convolutional neural network was used to extract features in the spatial domain and BLSTM was used to model contextual dependence. The experimental results showed that WA and UA reached 65.2% and 68.0% respectively.

The inputs of these deep learning models are mostly raw audio data and spectrograms extracted from audio signals, and a small part is manual features and high-level feature representations learned from various autoencoders. The end-to-end speech emotion recognition model is shown in Fig. 4.

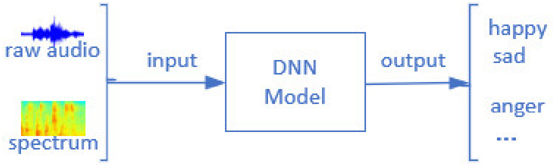

Fig. 4. Speech emotion recognition in an end-to-end way

Combining traditional machine learning methods with deep neural networks is increasingly becoming a trend. The framework of deep neural network hidden markov models called DNN-HMM was proposed in [40], and the best recognition accuracy is 53.89%. Zheng et al. [41] proposed the architecture of CNN-RF. The salient feature representation is extracted from CNN which acts as a feature extractor. Then the salient feature representation serves as the input of the random forest classifier. Experiments show that the classification accuracy of their proposed method is 3.25% higher than using only CNN.

IV. EXISTING PROBLEMS AND DIRECTION OF EXPLORATION

At present, the research of speech emotion recognition has achieved fruitful results. However, there are still many problems to be tackled urgently. Considering that most speech emotion recognition algorithms are supervised, the scarcity and insufficiency of labeled training data makes the performance of the model weak. Marking large amounts of data directly by humans is time-consuming and not an efficient solution. Some data enhancement methods can be adopted to expand the amount of data. Based on the Retinal Imaging Principle, each emotional utterance generates multiple spectrograms, which effectively expands the training set [35]. Some researchers try to use semi-supervised algorithms. In [9], in order to effectively use unlabeled data, the authors used a semi-supervised collaborative training algorithm. Some scholars have adopted the idea of transfer learning. In the future, more new data augmentation methods are worth discovering and exploring.

Generally, different people may have different understandings of the same emotional utterance, and the labeled data is subjective, because there is no uniform standard. In the future, the theoretical basis of emotion is worthy of further exploration. In addition, the emotions classified in many studies only involve very simple ones. In [36], the author only classified four basic emotions, neutral, angry, sad and happy. There is less research on complex emotions such as disgust and anxiety. In the future, more attention needs to be paid to complex emotion recognition.

From a model point of view, there is still no consensus regarding the best acoustic features in machine learning methods. More feature selection and feature extraction methods need to be discovered to extract representative and compact feature representations. One of the main problems with deep learning is that the model has poor explainability. These two aspects require a lot of effort, and breakthroughs in these two aspects may be milestone events in the field of speech emotion recognition.

V. CONCLUSION

Although, there are still many problems in the field of speech emotion recognition, which need to be further explored. It is reasonable to believe that the future of speech emotion recognition is bright. Intelligent robots with speech emotion recognition function will play an important role in the harmonious and natural human-robot interaction (HRI).

As an important part of affective computing, speech emotion recognition has been widely concerned, and many researchers have done a lot of research work. In this paper, we systematically sort out some mainstream models. It can be seen that researchers show more interest in deep learning in which features are mostly extracted automatically than the traditional machine learning methods. According to the time line, the previous work is summarized and discussed, and some basic conclusions are drawn. Some problems in the field of speech emotion recognition have been proposed, which need further research and exploration. At the same time, some parts of the summary are not comprehensive enough and need to be improved in the future.

REFERENCES

[1] J.-S. Park, J.-H. Kim, and Y.-H. Oh, “Feature vector classification based speech emotion recognition for service robots,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590–1596, 2009.

[2] Y. Li, C. T. Ishi, N. Ward, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, “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)*, pp. 1356–1359, IEEE, 2017.

[3] R. W. Picard, *Affective computing*. MIT press, 2000.

[4] M.-P. Jansen, “Communicative signals and social contextual factors in multimodal affect recognition,” in *2019 International Conference on Multimodal Interaction*, pp. 468–472, 2019.

[5] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[6] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

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

[8] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

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

[10] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 216–221, IEEE, 2013.

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

[12] C.-C. Tsai, Y.-Z. Chen, and C.-W. Liao, "Interactive emotion recognition using support vector machine for human-robot interaction," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, pp. 407–412, IEEE, 2009.

[13] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," in *Twelfth annual conference of the international speech communication association*, 2011.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

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

[17] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–413, IEEE, 2007.

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

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

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

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

[22] W. Zheng, J. Yu, and Y. 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)*, pp. 827–831, IEEE, 2015.

[23] J. J. Lasiman and D. P. Lestari, "Speech emotion recognition for indonesian language using long short-term memory," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 40–43, IEEE, 2018.

[24] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech.," in *Interspeech*, pp. 495–499, 2016.

[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "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)*, pp. 1–8, IEEE, 2013.

[26] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[28] Y. Kim and E. M. Provost, "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*, pp. 3677–3681, IEEE, 2013.

[29] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on affective computing*, vol. 6, no. 1, pp. 69–75, 2015.

[30] P.-Y. Shih, C.-P. Chen, and C.-H. Wu, "Speech emotion recognition with ensemble learning methods," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2756–2760, IEEE, 2017.

[31] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

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

[33] P. Harár, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 137–140, IEEE, 2017.

[34] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, pp. 1–5, IEEE, 2017.

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

[36] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition.," in *Interspeech*, pp. 3603–3607, 2016.

[37] 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)*, pp. 2227–2231, IEEE, 2017.

[38] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "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)*, pp. 5200–5204, IEEE, 2016.

[39] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "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*, pp. 27–33, 2018.

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

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