Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 3662839, 13 pages
https://doi.org/10.1155/2023/3662839

Research Article

Multimodal and Multitask Learning with Additive Angular Penalty Focus Loss for Speech Emotion Recognition

Guihua Wen,[1] **Sheng Ye**,[1] **Huihui Li**,[2] **Pengcheng Wen**,[1] **and Yuhan Zhang**[3]

[1]*School of Computer Science and Engineering, South China University of Technology, Guangzhou, China*
[2]*School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China*
[3]*Department of Neurology, Dongguan Songshanhu Central Hospital, Dongguan, China*

Correspondence should be addressed to Huihui Li; lihh@gpnu.edu.cn and Pengcheng Wen; 583283648@qq.com

Received 21 November 2022; Revised 11 March 2023; Accepted 3 October 2023; Published 17 October 2023

Academic Editor: Mohammad R. Khosravi

Speech emotion recognition has lots of applications such as human-computer interaction and health management. The current methods are challenged with the problems of fuzzy decision boundary and imbalance between difficult and easy samples in the training data. This paper first proposes an additive angle penalty focus loss function (APFL), which strictly refines the fuzzy decision boundary by introducing angle penalty factors to improve the compactness within the class and enlarge the distance between classes. It also assigns the larger loss to difficult samples to make the model pay more attention to them, as they are easily misclassified. Simultaneously, due to the lack of training samples, the framework of multimodal and multitask learning with APFL is further proposed, which extracts spectrogram features by deep neural network, text features by the pretrained language model, and audio features by the pretrained sound model. It uses the gender recognition as an auxiliary task. The experimental results verify the effectiveness of the proposed loss function and framework.

1. Introduction

Speech not only explicitly expresses linguistic content but also implicitly contains the speaker's emotional states such as sadness, happiness, and fear. The speech emotion recognition (SER) aims to automatically identify the speaker's emotional states [1–3], having a large number of applications such as human-computer interaction, information recommendation, and health detection. Consequently, methods for SER are deeply investigated. In addition to methods based on hand-crafted features [4, 5], some methods are mainly based on deep learning [6], which convert speech signals into spectrograms and then various deep learning methods are used to deal with them. Among them, deep convolutional neural networks (DCNNs) and recurrent neural networks are most widely used [7, 8]. The temporal convolutional networks are also popular for solving SER problem [9, 10]. They focus on the innovative design of the neural network structures [11, 12], such as by adding the attention mechanism [13] and the transfer learning to solve problems of SER [14].

The first problem is from the small training data [15], as the model trained on the small data will easily lead to the over fitting and in turn result in the weaker generalization ability. Data augmentation is an effective method to solve this problem [16]. For example, generative adaptive networks (GANs) and its variants are often applied to generate new samples [17–19]. Alternatively, a larger data can be directly constructed from existing data with hand-crafted features [20]. Because there is a large amount of unlabeled data, transfer learning [15] and semisupervised methods [21] can be also applied to expand the training data.

Another problem comes from insufficient features for SER. In such case, multimodal learning can be applied to learn enough features from different angles. These features are complementary so that they can describe the internal semantics of speech more completely and accurately. For example, speech and text can be integrated to extract features for SER [22–24]. In addition to speech and text, another method also considers facial expression and motion through

transformer encoder and then combines these features to achieve the classification [25]. These methods have achieved good results.

To solve the overfitting problem, multitask learning can be applied for SER. For example, it takes the language classification as an auxiliary task and speech emotion recognition as the main task [26]. It selects speech emotional features for any two classes independent of the speaker for the classification as multiple tasks and then ensembles their results [27]. The hierarchical multitask learning is proposed that uses the coarse classification and fine classification as two tasks [28]. There methods use unsupervised reconstruction as an auxiliary task [29]. The more complicated method obtains the multiscale unified metric [30], where the phone recognition and gender recognition are the auxiliary tasks. These methods are based on the single modal of speech.

The emotional labels of speech may be uncertain [1] for those methods based on frames for SER. When they segment each speech sentence into frames, the label of each frame is assigned by that of the sentence, easily leading to the noise labels [2]. Some new methods are proposed to solve this problem, such as the iterative self-learning framework that designed four specific label change rules [31] and the self-labeling method for each speech frame [2]. Simultaneously, the multiclassifier mutual learning is also proposed [1], where all classifiers classify each sample and then combine their classification results to construct its new label.

Some special issues are emphasized, such as the uneven length of input speech [32], which can be solved by DCNN and LSTM (long short-term memory). Besides, hand-crafted features of multivariable time series, bidirectional echo state network, and sampling methods are used to solve the unbalanced problem [33]. The individual standardization network aims to reduce the emotional confusion caused by individual differences [34]. To extract and select optimal features, the cryptographic structure [35], sparse coding [36], and the hybrid network of capsule network and transfer learning-based mixed task net are proposed [3]. In addition, ensemble deep learning [37] and supervised contrast learning [38] are also proposed.

DCNN needs loss function to guide its learning for SER. The most of loss functions are not specifically designed for speech emotion recognition [39, 40]. Although there are multimodal learning and multitask learning independently used for speech emotion recognition, they have not been combined. This paper proposes a new method that combines multimodal and multitask learning with new additive angular penalty focus loss (MTAP) to recognize the speech emotion. The main contributions are as follows:

(1) To solve the problem of the fuzzy decision boundary and the imbalance between difficult and easy speech samples, a new additive penalty focal loss function (APFL) is proposed for SER

(2) A new method is proposed for SER that combines multimodal and multitask learning with APFL, where the gender recognition is taken as an auxiliary task, and spectrogram, text, and audio are the different modalities of speech samples

Section 2 provides the related work, while the proposed method is introduced in Section 3. Experiment results and analysis are presented in Section 4. Section 5 presents conclusions.

2. Related Work

As our contributions are related to the combination of our new loss function with multimodal and multitask learning for speech emotion recognition, they are compared and analyzed.

2.1. Loss Function. The loss function widely used in speech emotion recognition is the cross-entropy loss (CEL) [2]. The center loss function [41] is also used for SER to pull features in the same emotional category to its center [40]. However, it only improves the intraclass compactness without enlarging the distance between classes. The triplet loss function [42] is also used for SER [39], which aims to reduce the distance of samples in the same class and enlarge the distance of heterogeneous samples. Another class-specific angular Softmax loss is designed to train the time-frequency convolution neural network [43]. In other fields, some new loss functions are also proposed, such as face recognition loss ArcFace [44] that transforms Euclidean space into the angle space and introduces additional angle penalties to target categories for strictly controlling the boundary of each category. This loss can achieve the effect of reducing the intraclass distance and increasing the interclass distance. Focal loss (FL) function [45] is proposed to solve the extreme imbalance between the foreground and background of data. It adjusts the contribution of hard samples to the total loss by introducing modulation parameters. These loss have not been used for SER. Particularly, different from these methods, our method combines ArcFace and FL in innovative way to solve the problem of fuzzy decision boundary and imbalance between difficult and easy samples. Generally, the deep neural network determines the gradient through the loss function and then uses it to modify the weights of the network. Our method works in the same way. But GHM (gradient harmonizing mechanism) [46] is different. Inspired by the gradient norm distribution, it first calculates the gradient density and then adds a harmonic parameter to the gradient of each sample according to the density. In practical applications, the modification of gradient can be realized equivalently by reconstructing the loss function. GHM changes with the density that may change in the training process. However, our method is a static loss function. It does not adapt to the change of data distribution. It also does

not change in the training process. However, GHM is currently used for the target detection, not for SER. Its principle can also be introduced into our method to further improve the performance.

2.2. Multimodal Learning. Multimodal learning can learn features from different modalities of samples. These features can be complementary so that they can describe the semantics of emotional speech more completely and accurately. For example, the method of integrating speech and text modality is proposed that emphasizes the temporal relationship between different modalities [22]. Another method also uses speech and text but introduces the attention network to promote the interaction and information fusion between them [23]. Alternatively, two different neural networks are applied to extract features of speech and text, respectively, and then concatenate their features [24]. These methods do not consider the relationship between different modalities. In addition to speech and text, other modalities such as facial expression and motion are considered, where the similarity between speech and text, and speech and other modal features are learned through transformer encoder and then their features are combined [25]. Different from these methods, our method combines the spectrogram features extracted by deep neural network, text features extracted by the pretrained language model BERT, and audio features extracted by the pretrained VGGish sound model.

2.3. Multitask Learning. Language and gender can affect the performance of speech emotion recognition [26], which can be used within the framework of multitask learning. For example, emotion classification is the main task and language classification is taken as an auxiliary task [26]. Taking gender recognition as an auxiliary task is conducive to extracting distinctive features and increasing the distinguishability between emotional categories [47–49]. Speakers are also used in multitask learning framework for speech emotion recognition, where each task selects features for any two classes independent of the speaker for the classification and then ensembles the classification results [27]. The hierarchical multitask learning framework is also proposed that takes the coarse classification and fine classification as two tasks [28]. The augmentation of data and unsupervised reconstruction can be taken as an auxiliary task to avoid the difficulties caused by the data annotation [29]. Another method is more complicated that obtains the multiscale unified metric [30] by the multitask learning, where the classification of both Emission States Category and Emission Intensity Scale is the main task and the classification of phone recognition and gender recognition is the auxiliary task.

There is one method that combines multimodel learning and multitask learning [50]. However, it aims to perform the speech recognition and identity recognition, instead of the speech emotion recognition, resulting in learned features that may deviate from the emotion recognition. Furthermore, it uses video, text, and audio. However, it is difficult to prepare the video data, as the speech sentences in video are not easy to be determined.

3. Additive Angle Penalty Focal Loss

This section proposes a new additive penalty focal loss function. Although ArcFace loss [44] and Focal loss [45] have been proposed in computer vision fields, they have not been considered in SER. Furthermore, both ArcFace and focal loss only consider one aspect of optimization such as fuzzy decision boundary or class imbalance, resulting in the limited improvements. Thus, we combine them in the innovative way to extract better discriminative features. It not only enhances intraclass compactness and interclass distance but also assigns more appropriate weights to the hard examples, so that it is obviously stronger in learning discriminative features than both ArcFace and focal loss. As it does not use the domain knowledge of SER, generally it can be also applied to other domains. In our case, we apply it to improve SER.

3.1. Additive Angle Penalty Focal Loss. Fuzzy decision boundary and class imbalance are the two challenges faced by speech emotion recognition. To tackle these issues and improve the recognition accuracy, APFL is devised as follows:

$$L_{\text{apfl}} = -\frac{1}{N}\sum_{i=1}^{N}\left(1 - p_{x_i,y_i}\right)^{\gamma}\log p_{x_i,y_i}, \tag{1}$$

where

$$p_{x_i,y_i} = \frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}, \tag{2}$$

and $\cos\theta_j = W_j^T x_i$. W_j denotes the j-th column of the weight matrix W after L2 normalization, x_i is the L2 normalized feature vector of the i-th sample corresponding to the ground-truth class y_i, θ_j is the angle distance between W_j and x_i. p_{x_i,y_i} denotes the posterior probability of x_i being classified to the class y_i. N is the number of training samples and n is the number of classes. s is a hyperparameter that should be adjusted carefully to obtain the optimal performance of the model. m is the additive penalty to the angle between x_i and the weight W_{y_i} of its corresponding label y_i so as to provide additional guidance to synchronously enhance the intraclass similarity and interclass difference. In this way, the issue of fuzzy decision boundary can be alleviated. γ is used to guide the model to pay more attention to the hard examples by multiplying the modulating factor $(1 - p_{x_i,y_i})^{\gamma}$.

The idea of APFL is quite useful and easy to implement in any deep-learning framework. Whenever Softmax loss or similar loss function is used, we can replace it with APFL to achieve the better performance.

3.2. Comparison with Different Loss Functions. In this section, we compare APFL with some relevant loss functions, i.e., Softmax loss (cross-entropy loss), focal loss, and ArcFace. For simplicity of analysis, we consider the binary classification case with classes C_1 and C_2.

3.2.1. Geometric Difference. As illustrated in Figure 1, we compare the decision boundaries. Evidently, APFL has stricter decision conditions (for C_1 it requires $\theta_1 \le \theta_2 - m$ rather than $\theta_1 \le \theta_2$, and it is similar for C_2), resulting in a clearer boundary with a margin of $\sqrt{2}m$ between different classes in the angular space.

3.2.2. Impact from Examples. Either Softmax loss or ArcFace does not take the influence of data distribution into account. Multiplying the modulation factor $(1 - p_{x_i,y_i})^{\gamma}$ alleviates this issue to an extent. Specifically, for examples that are easily misclassified, the factor is close to 1 as p_{x_i,y_i} is small; hence, the loss is nearly unaffected. But for those that are well classified, the factor goes to 0 since p_{x_i,y_i} tends to 1; thus, the loss is down-weighted. It can prevent the model from overwhelmed cases by easy examples. It can be easily found that these loss functions are in fact special cases of the proposed APFL. When $\gamma = 0$, APFL is equivalent to ArcFace. When $m = 0$, it becomes the focal loss with L2 normalized features and weights.

4. Multimodal and Multitask Learning Framework with APFL

This section introduces our proposed multimodal and multitask learning framework with our new loss (MTAP) for speech emotion recognition, shown in Figure 2, which uses spectrogram, text, and audio for the input speech sample while the gender recognition is taken as the auxiliary task.

4.1. Loss. Due to its effectiveness in speech emotion recognition [14], DCNN is used to extract features for the classifier to recognize the speech emotion. The input speech signal is first converted into the spectrogram and then feed it into DCNN to extract features. As illustrated in Figure 3, the dot product between extracted features and the last fully connected layer is equivalent to the cosine distance when both of them are normalized, where W is the weight matrix of the full connection layer and updated by the back-propagation of errors method. We use the arccos function to acquire the angle between them. Afterwards, we add an additive penalty to the angle and obtain the target logit back by using the cosine function. Subsequently, we rescale all logits by the fixed feature norm, and then following steps are exactly the same as in the Softmax loss. Finally, we multiply the cross-entropy by the modulating factor to adjust the less loss to the well classified examples.

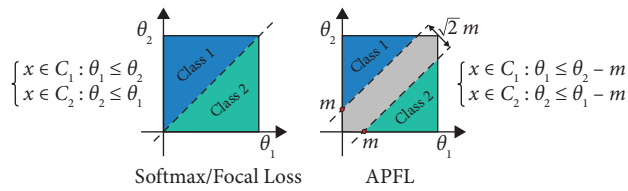

FIGURE 1: The comparison of decision margins of different loss functions under binary classification case. The dashed line represents the decision boundary, and the gray areas are decision margins.

Thus, the total loss for our framework is defined by $L_{\text{total}} = L_{\text{apfl}} + \lambda \times L_{\text{gender}}$, where L_{gender} for auxiliary task is defined by Softmax loss, L_{apfl} is for speech emotion recognition, and λ controls the influence of auxiliary task on the model.

4.2. BERT. BERT (Bidirectional Encoder Representations from Transformers) [51] is used to extract features of texts, which is a pretrained model. By combining tasks of both Masked Language Model (MLM) and Next Sentence Prediction (NSP), the embedded feature representation of language is learned by the self-supervised learning on a large corpus and then obtained features can be directly used as the input of downstream tasks. BERT has three parts as the input: Token embedding, Segment embedding, and Position embedding. Token embedding is the feature representation of the input text where Token can be understood as a word in Chinese. Segment embedding proceeds the paired sentences as the input, which has only two values: 0 and 1. For the input sentence pair, all Tokens of the previous sentence are given 0, and all Tokens of the next sentence are given 1. As text is sequential, the order of words will affect the meaning of sentences. However, Transformer structure cannot capture this information. Position embedding is designed to make up for this defect, which is learnable. Due to the complexity, the simpler version of BERT denoted as $\text{BERT}_{\text{BASE}}$ is used to extract features of the text corresponding to the speech.

4.3. VGGish. VGGish is a small VGG network [52], which was trained on a larger dataset AudioSet [53], which contains about 2.1 million videos with 527 sound categories. VGGish is the pretrained model that can be used as an audio feature extractor. It presents 128-dimensional feature vector for the input audio with one second, which can be used as the initial input of another model. Although there are differences between general audio and emotional speech, speech also contains some features of the general audio. These features can be also applied for SER [2]. As DCNN cannot learn enough features from the small emotional speech data sets, VGGish can be used to extract audio features as the complementary features. For each speech

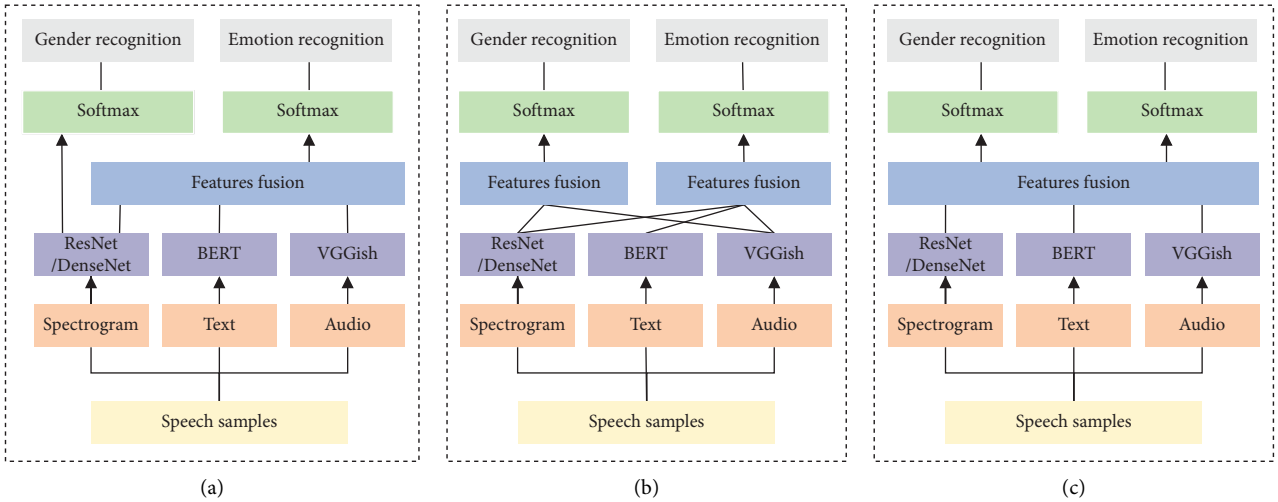

Figure 2: Framework of our MTAP for speech emotion recognition, where it fuses multimodal features and uses our new additive angle penalty focus loss function. It also considers the gender recognition as the auxiliary task in three different ways. MTL-A only uses spectrogram features for the auxiliary task, MTL-B uses both spectrogram and audio features, and MTL-C uses all features. (a) MTL-A. (b) MTL-B. (c) MTL-C.

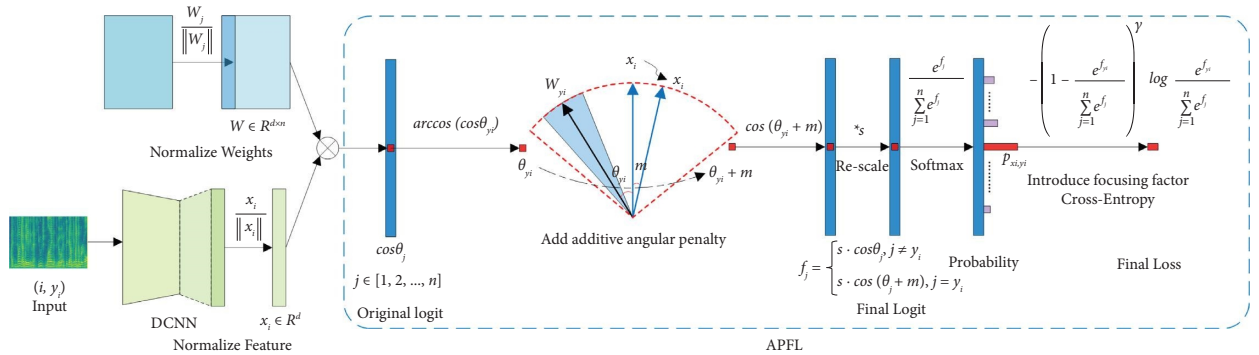

Figure 3: Framework of DCNN with new APFL loss for speech emotion recognition, where feature x_i and weight W_j are L2 normalized, and logit $\cos\theta_j$ is computed for each class. For the ground θ_{y_j}, an extra angular penalty m is added to calculate $\cos(\theta_{y_j} + m)$ as the new logit. All logits are then multiplied by the scaling parameter s and go through the Softmax to obtain the prediction probability p_{x_i,y_i} for computing the total loss.

sample, we use $\text{BERT}_{\text{BASE}}$ to extract its text features (F_{text}), use VGGish to extract its audio features (F_{audio}), use DCNN to extract its spectrogram features (F_{image}), and then fuse these features to obtain the comprehensive features $F = \text{Concat}(F_{\text{image}}, F_{\text{text}}, F_{\text{audio}})$ for the classifier to perform speech emotion recognition.

5. Experiments and Results

5.1. Datasets and Evaluation Indicators. Three benchmark databases are selected to evaluate our method, including Interactive Emotional Dyadic Motion Capture (IEMOCAP) [54], Surrey Audio-Visual Expressed Emotion (SAVEE) [55], and Berlin Emotional Speech Database (EMODB) [56].

IEMOCAP [54] consists of 5 sessions and each session is displayed by a pair of speakers (male and female) in scripted and improvised scenarios. We choose 4 emotion types (i.e., angry, happy, neutral, and sad) for our experiments only from improvised data, and thus, 2280 utterances are used. We adopt 5-fold cross-validation using Leave-One-Session-Out (LOSO) strategy, that is, 4 sessions are used for training, while the rest one is divided into two equal parts for validation and testing.

SAVEE [55] is composed of records performed by four native English male actors in seven emotions. It includes 480 utterances in total, i.e., 60 anger, 60 disgust, 60 fear, 60 happiness, 120 neutral, 60 sadness, and 60 surprise. On these data, we conduct fourfold cross-validation Speaker Independent (SI) and fivefold cross-validation speaker dependent (SD) experiments.

EMODB [56] contains 535 emotional utterances performed by 10 actors with seven different emotions: anger, boredom, disgust, fear, happiness, neutral, and sadness. On these data, we adopt 10-fold cross-validation strategy for both SI and SD experiments.

Generally, the performance of SER can be evaluated by two widely used metrics [57–59]. One is the *Weighted Accuracy (WA)* that is the classification accuracy of all test

samples, also known as Overall Accuracy. The other is *Unweighted Accuracy (UA)* that is the average accuracy of each individual class, also known as Class Accuracy.

5.2. Experimental Results of Our Proposed New Loss Function

5.2.1. Data Preprocessing. We trim the long duration audio utterances to shorter duration ones which covers 75 percentiles of all audio samples in IEMOCAP [57]. Thus, the maximum duration is restricted to six seconds. For those longer than six seconds, their head and tail are cut. Each trimmed sample is assigned the same emotion label as that of its utterance. Subsequently, we use the feature extraction method [58] to obtain spectrograms, where the sequence of overlapping hamming windows is applied with the frame length of 40 msec and frame interval of 10 msec. For each frame, we calculate its discrete Fourier transform and then perform the aggregation of the short-time spectra to obtain a matrix of size $T \times F$, where $T \leq 600$ and $F = 400$. The last step uses the zero padding to obtain the fixed 600 time points. Thus, the spectrogram size is 600×400 for IEMOCAP, 500×400 for SAVEE, and 400×400 for EMODB.

5.2.2. Experimental Settings. DenseNet169 [60] pretrained on ImageNet [61] is used to extract features of speech spectrograms. The parameter $\gamma \in \{0.10, 0.20, 0.50, 1.0, 2.0, 5.0, \}$, the penalty $m \in [0.2, 0.5]$, and the feature scale s is an empirical parameter that should be appropriately large where it equals to 10. All experiments use the cross-validation strategy. Besides, we run five times per fold and then take the average as the final result of the fold to ensure the reliability.

5.2.3. Speaker Independent Experiments. Under this strategy, we conduct experiments on both EMODB and SAVEE using 10-fold and 4-fold cross-validation, respectively. The results are reported using the average value and standard deviation of WA and UA. It can be found from Table 1 that our new APFL outperforms all compared methods in WA and UA. In addition, its standard deviation is also minimum in most cases, indicating that it can make the model more stable.

5.2.4. Speaker Dependent Experiments. The relevant experimental results are reported in Table 2. It can be seen that our new APFL still obviously outperforms all compared methods in WA and UA. It also achieves the more significant improvements than that in Speaker Independent cases. Especially, compared with ArcFace loss on EMODB, APFL has the larger improvements of nearly 2% in terms of both WA and UA. Similarly, it can be observed that the model using APFL is more stable on the whole in terms of standard deviation.

5.2.5. Leave-One-Session-Out Experiments. As described earlier, we choose the improvised speech part from IEMOCAP as it comes from real cases. Experiments are conducted by LOSO with fivefold cross-validation. The optimal parameters and results are reported in Table 3 in the format of means and the standard deviations of WA and UA. It can be found that the model with APFL performs best among all models with compared loss functions.

5.2.6. Visual Analysis of Loss Functions. In order to illustrate advantages of our new loss function, we use t-SNE (t-distribution stochastic neighbor embedding) method [62] to visualize the distribution of features extracted from DenseNet169 on test samples in IEMOCAP under the guidance of each compared loss function. The results are shown in Figure 4.

It can be seen in European space that the category decision boundary of CE Loss is very vague, the categories are basically mixed together, and the overall distribution is very loose. Although focal loss performs slightly better, it is still messy. ArcFace has a significant improvement, having three distinct clusters. However, happy category in yellow is basically confused with the other three categories. In contrast, clusters formed by our APFL are banded clearly in four directions, corresponding to four categories. In particular, the boundary between the happy category in yellow and the angry category in blue becomes clearer, which indicates that some samples ever wrongly classified by the other loss functions have now been correctly identified. The similar results can be observed in the angle space. It can be seen that the category boundary of CE loss is much vague, basically mixed. In such a case, it is hard to perform the nice classification. Although focal loss obtains the better results, its boundaries among classes are still overlapped. By comparison, ArcFace seems form three separated clusters. However, four clusters are still not clearly formed. In contrast, clusters formed by our APFL are separated in four segments, corresponding to four categories. The interval between categories is also obviously larger than that of the other loss functions, while the arc length in the same category also becomes smaller. This means that the intraclass compactness and interclass difference by our APFL have been improved.

5.3. Experimental Results of Our MTAP.

As our MTAP consists of several components, ablation experiments are conducted to illustrate the necessity and superiority of each component.

5.3.1. Multimodal. In order to more clearly understand the effectiveness of each modal, ablation experiments are conducted on the improvised part of IEMOCAP. The results are shown in Table 4. It can be seen that all modalities are necessary to improve the performance. MTAP achieves an obvious improvement of 2% on UA when the text modal is added to spectrogram modal. However, the further improvement is limited when the audio modal is further added, indicating that there may be redundancy between three modalities. In order to more intuitively reflect the contribution of each modal to features learned by MTAP,

Table 1: Experimental results by SI, where DenseNet169 is used to extract features for the spectrogram.

Database	Loss	γ	m	WA	UA
EMODB	Softmax	—	—	74.37 ± 0.55	69.74 ± 0.52
	Focal [45]	0.10	—	74.81 ± 0.63	69.78 ± 0.59
	ArcFace [44]	—	0.50	80.66 ± 0.19	77.88 ± 0.92
	APFL (ours)	0.10	0.50	**81.09 ± 0.27**	**78.73 ± 0.54**
SAVEE	Softmax	—	—	49.24 ± 0.94	43.53 ± 0.90
	Focal [45]	0.10	—	49.38 ± 0.74	43.81 ± 1.02
	ArcFace [44]	—	0.50	52.85 ± 0.60	48.21 ± 1.22
	APFL (ours)	5.0	0.50	**53.33 ± 0.85**	**49.13 ± 0.50**

Bold fonts indicate the best performance.

Table 2: Experimental results by SD, where DenseNet169 is used to extract features for the spectrogram.

Database	Loss	γ	m	WA	UA
EMODB	Softmax	—	—	78.32 ± 0.82	76.11 ± 0.88
	Focal [45]	0.10	—	79.25 ± 0.46	77.02 ± 0.86
	ArcFace [44]	—	0.50	84.85 ± 0.69	82.93 ± 0.78
	APFL (ours)	5.0	0.50	**86.42 ± 0.49**	**84.77 ± 0.50**
SAVEE	Softmax	—	—	61.81 ± 1.77	57.34 ± 2.13
	Focal [45]	0.10	—	62.08 ± 0.17	57.66 ± 0.15
	ArcFace [44]	—	0.50	66.39 ± 0.56	63.17 ± 0.43
	APFL (ours)	0.50	0.50	**67.15 ± 0.20**	**63.65 ± 0.59**

Bold fonts indicate the best performance.

Table 3: Experimental results by LOSO, where DenseNet169 is used to extract features for the spectrogram.

Database	Loss	γ	m	WA	UA
IEMOCAP	Softmax	—	—	71.56 ± 0.79	60.61 ± 1.20
	Focal [45]	0.50	—	71.89 ± 0.46	61.94 ± 1.16
	ArcFace [44]	—	0.30	72.48 ± 1.24	63.67 ± 1.02
	APFL (ours)	0.50	0.30	**72.83 ± 0.54**	**64.78 ± 0.93**

Bold fonts indicate the best performance.

we use t-SNE to visualize the distribution of test samples whose features are extracted by our model in European space. The results are shown in Figure 5. When only the spectrogram features is used, it can be seen that there are roughly three clusters, but they are basically mixed together without clear boundaries. However, by comparison, our results in Figure 5(d) have the better gap among three clusters denoting anger, neutral, and sad, indicating that multimodal can improve the recognition performance of each emotion category, proving that our method is effective.

5.3.2. Gender Identification as Auxiliary Task. In the proposed MTAP, the gender recognition is taken as the auxiliary task. Experiments are conducted to illustrate its effectiveness. The results are shown in Table 5. It can be seen that the auxiliary task has improved the performance of the model in any case, illustrating that text modal is effective for SER. In more details, MTL-B performs best in three cases, even surpassing MTL-C, where their difference is only whether the text modal is used. This means that text modal is not effective for the gender recognition task. It is reasonable because features of the text do not vary with the gender difference but only related to the content of the text itself.

5.3.3. New Loss Function. To validate that APFL is the necessary component of MTAP, some experiments are conducted where both CEL and APFL as loss functions are used and all others remain unchanged for MTAP. The results are shown in Table 6. It can be seen that MTAP with ResNet50 has achieved 2% improvement in both WA and UA, while it with DenseNet-169 has achieved improvement by 1.5% in WA and by 2.7% in UA. The effects are much significant, proving the superiority and necessity of APFL to MTAP. By further combining it with multimodal information, MTAP has achieved the significant improvements of 3% and 5%, respectively, on WA and UA.

5.4. Comparison with Recent Methods. In order to verify the superiority of our MTAP, several advanced methods in recent years are applied to make comparison with MTAP, as these methods use the backbone network similar to ours and experimental settings are the same. The experimental data are the improvised part of IEMOCAP, as it is closer to real

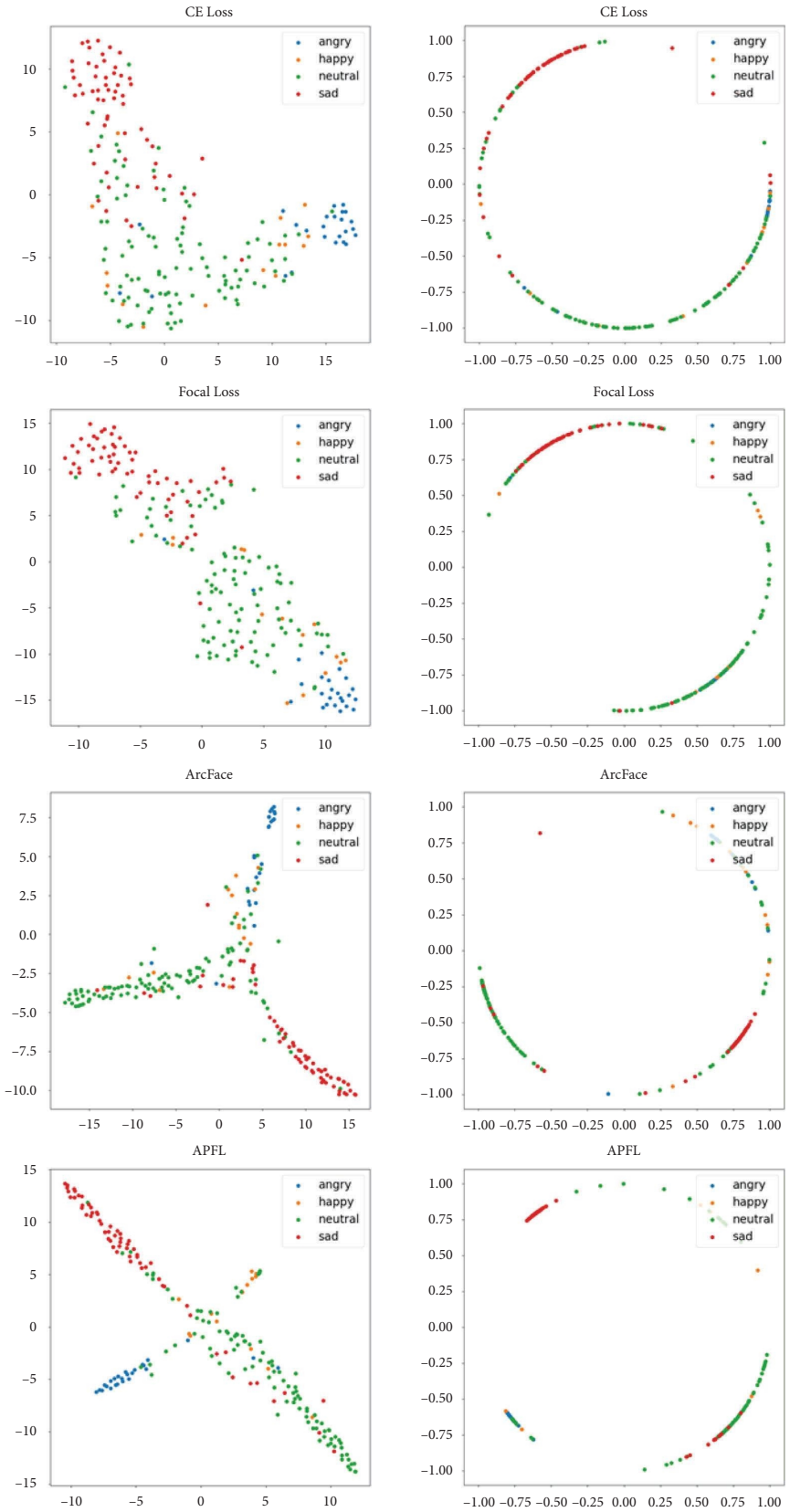

Figure 4: Visualization of samples distribution in IEMOCAP whose features are extracted by different loss functions, where the first and second columns, respectively, represent distributions in European space and angle space.

Table 4: Performance comparison of our MTAP in the case of different combinations of multiple modalities.

Models	Image	Text	Audio	WA (%)	UA (%)
ResNet50	√			70.82 ± 1.08	60.25 ± 2.34
+BERT	√	√		71.60 ± 0.97	61.25 ± 0.79
+VGGish	√		√	71.58 ± 1.03	60.48 ± 1.53
+BERT + VGGish	√	√	√	**71.84 ± 0.62**	**61.87 ± 2.26**
DenseNet169	√			71.56 ± 0.79	60.61 ± 1.20
+BERT	√	√		71.83 ± 0.82	62.59 ± 1.36
+VGGish	√		√	71.72 ± 0.89	61.41 ± 1.16
+BERT + VGGish	√	√	√	**71.85 ± 0.59**	**62.91 ± 0.75**

Bold fonts indicate the best performance.

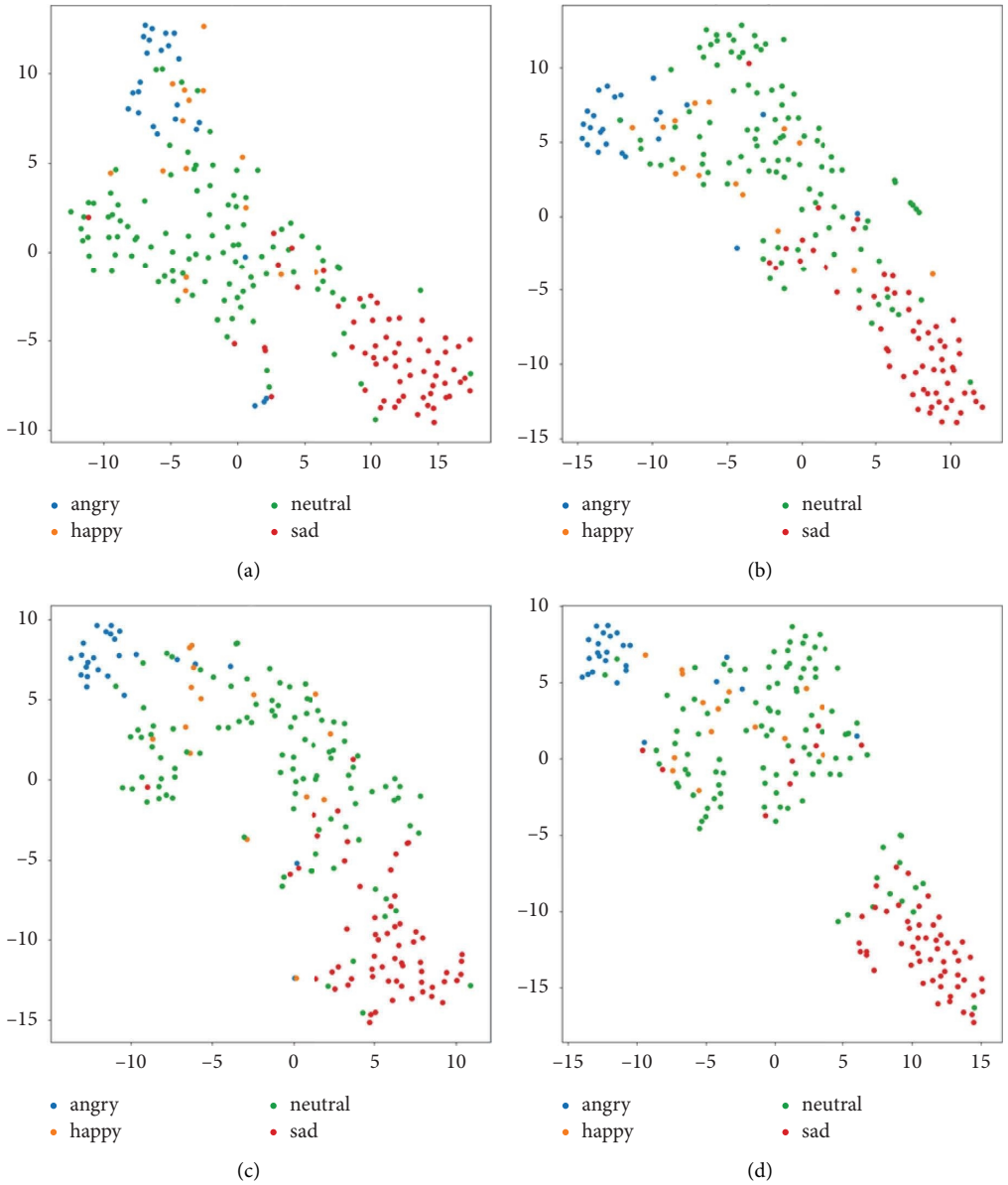

Figure 5: Visualization of test samples distribution in IEMOCAP whose features are extracted with the combination of different modalities, where (a) spectrogram, (b) spectrogram with text, (c) spectrogram with audio, and (d) spectrogram with text and audio.

Table 5: Performance comparison of our MTAP where gender identification is the auxiliary task.

Models	λ	WA (%)	UA (%)
ResNet50	0	71.84 ± 0.62	61.87 ± 2.26
+BERT + VGGish + MLT-A	0.25	72.32 ± 1.08	62.55 ± 0.59
+BERT + VGGish + MLT-B	0.25	**72.58 ± 0.37**	**62.93 ± 0.99**
+BERT + VGGish + MLT-C	0.25	71.94 ± 1.18	62.36 ± 1.09
DenseNet169	0	71.85 ± 0.59	62.91 ± 0.75
+BERT + VGGish + MLT-A	1.0	72.41 ± 1.14	63.02 ± 1.14
+BERT + VGGish + MLT-B	0.5	72.54 ± 1.05	**63.98 ± 1.79**
+BERT + VGGish + MLT-C	1.0	**73.07 ± 0.45**	63.08 ± 0.98

Bold fonts indicate the best performance.

Table 6: Performance comparison of MTAP in the case of different loss functions and backbone networks, where $m = 0.3$.

Models	Loss	WA (%)	UA (%)
ResNet50	CEL	71.84 ± 0.62	61.87 ± 2.26
+BERT + VGGish	CEL	71.84 ± 0.62	61.87 ± 2.26
ResNet50	APFL ($\gamma = 2.0$)	73.56 ± 1.12	62.60 ± 1.32
+BERT + VGGish	APFL ($\gamma = 2.0$)	**73.84 ± 0.75**	**63.75 ± 0.94**
DenseNet169	CEL	71.56 ± 0.79	60.61 ± 1.20
+BERT + VGGish	CEL	71.85 ± 0.59	62.91 ± 0.75
DenseNet169	APFL ($\gamma = 0.5$)	72.83 ± 0.54	64.78 ± 0.93
+BERT + VGGish	APFL ($\gamma = 0.1$)	**73.29 ± 0.73**	**65.64 ± 0.91**

Bold fonts indicate the best performance.

Table 7: Performance comparison of MTAP with recent methods on the improvised part of IEMOCAP.

Models	WA (%)	UA (%)
CNN-BLSTM [7]	68.80	59.40
CNN-BLSTM with a two-step predictor [7]	67.30	62.00
Parallel CNN [57]	71.20	61.90
CNN-GRU-SeqCaps [63]	72.73	59.71
Variable-length CNN-GRU [64]	71.45	64.22
CNN-TF-GAP [43]	72.43	64.80
End-to-end ASR and SER [65]	69.70	63.10
CNN-MHSA [59]	72.34	58.31
MTAP with ResNet50 (ours)	**73.84**	63.75
MTAP with DenseNet169 (ours)	73.29	**65.64**

Bold fonts indicate the best performance.

situations. These compared methods are basically developed from CNN structure, but they use only one or two modalities of speech. It seems that there is no method at present using three modalities of spectrogram, text, and audio. It can be seen from Table 7 that our method is optimal in both WA and UA, illustrating the superiority of our method.

6. Conclusions

This paper proposes a new multimodal and multitask learning method for speech emotion recognition, where a new additive angle penalty focus loss function is also proposed to guide the network learning. One of its advantages is that spectrogram features, text features, and audio features are extracted from different angles and then combined to enrich features for speech emotion recognition. Another advantage is that the auxiliary task of the gender identification is applied to improve the generalization ability and transfer the knowledge to SER for the complementary. This is because there are different properties of voice signals that male and female express the same emotion. When the neural network model is used to learn the emotional features of the input voice signal, if the gender is not specified, the model must learn emotional features composed of both male and female simultaneously, so that the feature space is not only large but also sparse. In such a case, a lot of training samples are required; otherwise it is easy to overfit. When the gender recognition task is introduced, the model can learn the emotional feature space of male and female, respectively, equal to learning two smaller subspaces. Generally, the dimension of the subspace is smaller than that of the whole space, so that it needs smaller training samples. In the case of the same training samples, the model with the gender

recognition task is more capable of learning ability, leading to the higher accuracy of speech emotion recognition. Furthermore, the proposed APFL has advantages of improving the compactness within the class, enlarging the difference between classes and focusing on difficult samples, so as to guide the network to learn more effective emotional features. Although our method has achieved the good performance, there is still room for the further improvement. For example, more modalities can be considered such as the speaker's facial expression and body movements, while the more complicated backbone network can be also considered such as with attention mechanism. They will be investigated in the future work, including their applications such as recognition of Parkinsons's disease through speech signals.

Data Availability

The data that support the findings of this study are available at https://sail.usc.edu/iemocap/iemocap_release.htm (IEMOCAP [54]), https://kahlan.eps.surrey.ac.uk/savee/Download.html (SAVEE [55]), and https://kahlan.eps.surrey.ac.uk/savee/Download.html (EMODB [56]).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by National Natural Science Foundation of China (Grant nos. 62176095 and 62006049), Guangzhou Science and Technology Planning Project (Grant nos. 2023B03J1336 and 2023A03J0316), Guangdong Province Key Area R&D Plan Project (Grant no. 2020B1111120001), Guangdong Basic and Applied Basic Research Foundation (Grant no. 2023A1515010939), and Project of Education Department of Guangdong Province (Grant nos. 2022KTSCX068 and 2021ZDZX1079).

References

[1] Y. Zhou, X. Liang, Y. Gu, Y. Yin, and L. Yao, "Multi-classifier interactive learning for ambiguous speech emotion recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 695–705, 2022.

[2] G. Wen, H. Liao, H. Li et al., "Self-labeling with feature transfer for speech emotion recognition," *Knowledge-Based Systems*, vol. 254, 2022.

[3] X.-C. Wen, J.-X. Ye, Y. Luo et al., "CTL-MTNet: A novel CapsNet and transfer learning-based mixed task net for the single-corpus and cross-corpus speech emotion recognition," in *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence*, pp. 2305–2311, Vienna, Austria, July 2022.

[4] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22–35, 2019.

[5] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden Markov models for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6715–6719, Brighton, UK, May 2019.

[6] A. Bakhshi, A. Harimi, and S. Chalup, "CyTex: transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[7] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 1089–1093, Stockholm, Sweden, August 2017.

[8] G. Lili, W. Longbiao, D. Jianwu, Z. Linjuan, G. Haotian, and L. Xiangang, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 1611–1615, Hyderabad, India, September 2018.

[9] J. X. Ye, X. C. Wen, X. Z. Wang et al., "GM-TCNet: gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition," *Speech Communication*, vol. 145, pp. 21–35, 2022.

[10] Y. Jiaxin, W. Xin-cheng, W. Yujie, X. Yong, L. Kunhong, and S. Hongming, "Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition," 2022, https://arxiv.org/abs/2211.08233.

[11] L. Runnan, W. Zhiyong, J. Jia, Z. Sheng, and M. Helen, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6675–6679, Brighton, UK, May 2019.

[12] L. Jiaxing, L. Zhilei, W. Longbiao, G. Lili, and D. Jianwu, "Speech emotion recognition with local-global aware deep representation learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7174–7178, Barcelona, Spain, May 2020.

[13] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[14] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[15] R. Li, J. Zhao, and Q. Jin, "Speech emotion recognition via multi-level cross-modal distillation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, pp. 606–610, Brno, Czechia, September 2021.

[16] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, 2022.

[17] A. Shilandari, H. Marvi, H. Khosravi, and W. Wang, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *Signal, Image and Video Processing*, vol. 16, no. 7, pp. 1955–1962, 2022.

[18] B. H. Su and C.-C. Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-GAN," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1991–2004, 2023.

[19] S. Wang, H. Hemati, J. Guenason, and D. Borth, "Generative data augmentation guided by triplet loss for speech emotion recognition," 2022, https://arxiv.org/abs/2208.04994.

[20] K. Manohar and E. Logashanmugam, "Hybrid deep learning with optimal feature selection for speech emotion recognition

using improved meta-heuristic algorithm," *Knowledge-Based Systems*, vol. 246, Article ID 108659, 2022.

[21] E. N. N. Ocquaye, Q. Mao, Y. Xue, and H. Song, "Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network," *International Journal of Intelligent Systems*, vol. 36, no. 1, pp. 53–71, 2021.

[22] G. N. Dong, C. M. Pun, and Z. Zhang, "Temporal relation inference network for multimodal speech emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6472–6485, 2022.

[23] Y. Tang, Y. Hu, L. He, and H. Huang, "A bimodal network based on Audio-Text-Interactional-Attention with ArcFace loss for speech emotion recognition," *Speech Communication*, vol. 143, pp. 21–32, 2022.

[24] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 4, pp. 2927–2931, Brno, Czechia, September 2021.

[25] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Computers and Industrial Engineering*, vol. 168, Article ID 108078, 2022.

[26] Z. Kexin and L. Yunxiang, *Speech Emotion Recognition System Based On Wavelet Transform And Multi-Task Learning*, SSRN, New, York, NY, USA, 2022.

[27] E. Kalhor and B. Bakhtiari, "Speaker independent feature selection for speech emotion recognition: a multi-task approach," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8127–8146, 2021.

[28] Z. Huijuan, Y. Ning, and W. Ruchuan, "Coarse-to-Fine speech emotion recognition based on multi-task learning," *Journal of Signal Processing Systems*, vol. 93, no. 2-3, pp. 299–308, 2021.

[29] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," 2022, https://arxiv.org/abs/2207.05298.

[30] Z. Kang, J. Peng, J. Wang, and J. Xiao, "SpeechEQ: Speech emotion recognition based on multi-scale unified datasets and multitask learning," 2022, https://arxiv.org/abs/2206.13101.

[31] S. Mao, P. C. Ching, and T. Lee, "Enhancing segment-based speech emotion recognition by iterative self-learning," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 123–134, 2022.

[32] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 680–688, 2022.

[33] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Bidirectional parallel echo state network for speech emotion recognition," *Neural Computing and Applications*, vol. 34, no. 20, pp. 17581–17599, 2022.

[34] W. Fan, X. Xu, B. Cai, and X. Xing, "ISNet: individual standardization network for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1803–1814, 2022.

[35] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems*, vol. 211, no. 9, Article ID 106547, 2021.

[36] P. Song, W. Zheng, Y. Yu, and S. Ou, "Speech emotion recognition based on robust discriminative sparse regression," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 2, pp. 343–353, 2021.

[37] R. Barkur, Deepanshi, D. Suresh, T. N. Mahesh Kumar, and A. V. Narasimhadhan, "EnsembleWave: an ensembled approach for automatic speech emotion recognition," in *Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies*, Bangalore, India, July 2022.

[38] V. S. Alaparthi, T. R. Pasam, D. A. Inagandla, J. Prakash, and P. K. Singh, "ScSer: supervised contrastive learning for speech emotion recognition using transformers," in *Proceedings of the International Conference on Human System Interaction, HSI*, Melbourne, Australia, July 2022.

[39] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 3673–3677, Hyderabad, India, September 2018.

[40] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7405–7409, Hyderabad, India, September 2019.

[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 499–515, Springer, Berlin, Germany, November 2016.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, Boston, MA, USA, June 2015.

[43] Z. Li, L. He, J. Li, L. Wang, and W.-Q. Zhang, "Towards discriminative representations and unbiased predictions: class-specific angular softmax for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 1696–1700, Graz, Austria, September 2019.

[44] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2022.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[46] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 8577–8584, 2019.

[47] H. Houari and M. Guerti, "Study the influence of gender and age in recognition of emotions from Algerian dialect speech," *Traitement du Signal*, vol. 37, no. 3, pp. 413–423, 2020.

[48] C. Gorrostieta, R. Lotfian, and K. Taylor, "Gender de-biasing in speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 2823–2827, Graz, Austria, September 2019.

[49] P. Vasuki and R. Divya Bharati, "Speech emotion recognition based on gender influence in emotional expression," *International Journal of Intelligent Information Technologies*, vol. 15, no. 4, pp. 22–40, 2019.

[50] A. Khare, S. Parthasarathy, and S. Sundaram, "Multi-Modal embeddings using multi-task learning for emotion recognition," in *Proceedings of the Interspeech*, pp. 384–388, October 2020.

[51] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1041–1049, Mountain View, CA, USA, October 2017.

[52] M. Abadi, A. Agarwal, P. Barham et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," 2016, https://arxiv.org/abs/1603.04467.

[53] S. Haq, P. J. B. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pp. 185–190, Australia, September 2008.

[54] C. Busso, M. Bulut, C.-C. Lee et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[55] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proceedings of the International Conference on Audio-Visual Speech Processing*, pp. 53–58, Norwich, UK, October 2009.

[56] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1517–1520, Lisbon, Portugal, September 2005.

[57] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram and phoneme embedding," in *Proceedings of the Interspeech*, pp. 3688–3692, Hyderabad, India, August 2018.

[58] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, pp. 1089–1093, Stockholm, Sweden, August 2017.

[59] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7189–7193, Barcelona, Spain, May 2020.

[60] G. Huang and Z. Liu, "Densely connected convolutional networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–4, Honolulu, HI, USA, July 2017.

[61] D. Jia, W. Dong, R. Socher, L. Jia, K. Li, and L. Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[62] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[63] X. Wu, Y. Cao, H. Lu et al., "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3280–3291, 2021.

[64] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 3683–3687, Hyderabad, India, November 2018.

[65] H. Feng, S. Ueno, and T. Kawahara, "End-to-End speech emotion recognition combined with acoustic-to-word ASR model," in *Proceedings of the Interspeech*, pp. 501–505, Organization of Virtual Conference Sessions, October 2020.