Speech Emotion Recognition via an Attentive Time–Frequency Neural Network

Cheng Lu, Wenming Zheng, Senior Member, IEEE, Hailun Lian, Yuan Zong, Member, IEEE, Chuangao Tang, Sunan Li, and Yan Zhao

Abstract—Spectrogram is commonly used as the input feature of deep neural networks to learn the high(er)-level time–frequency pattern of speech signal for speech emotion recognition (SER). Generally, different emotions correspond to specific energy activations both within frequency bands and time frames on spectrogram, which indicates the frequency and time domains are both essential to represent the emotion for SER. However, recent spectrogram-based works mainly focus on modeling the long-term dependency in time domain, which makes these methods suffer from the following issues: 1) neglecting to model the emotion-related correlations within frequency domain during the time–frequency joint learning and 2) ignoring to capture the specific frequency bands associated with emotions. To cope with the issues, we propose an attentive time–frequency neural network (ATFNN) for SER, including a time–frequency neural network (TFNN) and time–frequency attention. Specifically, aiming at the first issue, we design a TFNN with a frequency-domain encoder (F-Encoder) based on the Transformer encoder and a time-domain encoder (T-Encoder) based on the bidirectional long short-term memory (Bi-LSTM). The F-Encoder and T-Encoder model the correlations within frequency bands and time frames, respectively, and they are embedded into a time–frequency joint learning strategy to obtain the time–frequency patterns of speech emotions. Moreover, to handle the second issue, we adopt the time–frequency attention with a frequency-attention network (F-Attention) and a time-attention network (T-Attention) to focus on the emotion-related long-range dependencies between frequency bands and across time frames, which can enhance the emotional discrimination of speech features. Extensive experimental results on three public emotional databases, i.e., IEMOCAP, ABC, and CASIA, show that our proposed ATFNN outperforms the state-of-the-art methods.

Index Terms—Attention mechanism, spectrogram, speech emotion recognition (SER), time–frequency neural network (TFNN).

I. INTRODUCTION

SPEECH emotion recognition (SER) task is to make the machine automatically recognize the emotional states from the human speech. Recently, it has been a research hotspot in affective computing and pattern recognition [1], [2], [3], [4], [5]. Generally speaking, the key to deal with the SER problem is to extract the discriminative and generalized features of emotional speech [6], [7], [8], [9], [10]. These features mainly contain two categories at present: the handcrafted features and the deep features.

The hand-crafted features are primarily adopted in the earlier SER works, known as the low-level descriptors (LLDs) [11], [12], e.g., frame energy/loudness, fundamental frequency, zero/mean-crossing rate, and Mel-frequency-Cepstral coefficients (MFCC). A mount of related works explored various LLDs or these combinations to improve the performance of the SER model. For instance, Schuller et al. [11] integrated 6552 hand-crafted features into an open-source toolkit, i.e., openEAR, for four affective recognition tasks and then extended the openEAR to a multifunctional and convenient toolkit for diverse speech-related tasks, i.e., openSMILE [12]. In the earlier works, these LLDs and their combinations were widely utilized as the benchmark features for SER [6], [7].

Recently, with the emergence of deep learning, handcrafted features are gradually replaced by deep features extracted through deep neural networks (DNNs), e.g., deep convolutional neural networks (DCNNs) [8], [9] and recurrent neural networks (RNNs) [13], [14], [15]. Instead of LLDs, these deep features are high(er)-level features. In the extraction of deep features, generally, spectrogram features (e.g., MFCC, magnitude spectrogram, and Mel-spectrogram) with rich time–frequency information are often used as input features of DNNs to learn the high(er)-level time–frequency patterns of speech signal for SER [7], [8], [9], [13]. Mao et al. [8] investigated to utilize the dimensionality-reduced magnitude spectrogram to learn the salient features by CNNs. Zhang et al. [9] extended the spectrogram to a 3-D Mel-spectrogram to learn utterance-level speech emotion features from segment features by a discriminant temporal pyramid matching strategy. Moreover, Wang et al. [13] combined the MFCC features and Mel-spectrogram to input a special long short-term memory (LSTM), called dual-sequence LSTM, to extract the utterance-level emotion features.

The extraction of deep features is increasingly embedded in a time–frequency joint learning strategy to generate...
more effective time–frequency patterns of speech emotions [16], [17], [18], [19]. Satt et al. [18] proposed a combined convolution-LSTM network to extract the utterance-level feature with segment-level features. Chen et al. [19] also adopt the 3-D attention-based convolutional recurrent neural network (3-D ACRNN) to represent the utterance-level affective-salient features of emotional speech.

Nevertheless, the related works of the time–frequency feature extraction using spectrograms commonly encounter two issues at present. The first issue is that the existing works mainly focus on the correlations between speech frames to characterize the long-range dependency in time domain [8], [13], [17]. However, some works have revealed that the energy of the frequency band is changing in different emotions [20], [21], [22]. Cowie et al. [21] reported the variations of acoustic parameters associated with emotions, and the results of the acoustic spectral parameters indicated that each emotion usually corresponds to some energy activations of specific frequency band ranges on the spectrogram. For instance, the investigation in [21] shows that happiness has increased activations in high-frequency energy, while sadness has decreased activations in high-frequency energy. These situations indicate that capturing the energy variations in different frequency ranges is also essential for the representation of speech emotions. Therefore, this article will jointly model the correlations both within the frequency and time domains to obtain robust time–frequency representations of emotions.

Another issue faced by the time–frequency representation of speech emotion is that the distribution of emotion information in an utterance is sparse in both frequency and time domains [19], [23]. Specifically, it is intuitive that there are some frames without contents in a long utterance. Thus, not all frames in a sentence contain emotional information. In other words, the frames of an utterance have different contribution degrees in the representation of emotional speech, i.e., there are key frame regions highly associated with emotions. For the purpose of capturing these emotion-related regions, Chen et al. [19] investigated adding an attention layer after CNN and LSTM to capture specific temporal frames for the speech emotion representation. Also, Zhang et al. [23] adopted an attention mechanism to empower the sublayer to focus on emotion salient regions of the spectrogram.

However, these studies mainly embed attention into high-level time–frequency features, leading to introducing the redundant information in the feature extraction and roughly locating the emotion-related regions in the time–frequency domain. In addition, they mainly focus on the selection on the time frames while ignoring the selection on the frequency bands. Actually, some studies have proven that each emotion does not have energy activation on all frequency bands in the frequency domain [20], [21], [22]. Therefore, we need to capture the key time frames in the time domain, and the key frequency bands in the frequency domain, which are all highly related to emotions.

To deal with the issues mentioned above, in this article, we propose an attentive time–frequency neural network (ATFNN) to learn the discriminative speech emotion feature for SER, as shown in Fig. 1. In detail, we propose a time–frequency neural network (TFNN), including a frequency-domain encoder (F-Encoder) to model frequency features and a time-domain encoder (T-Encoder) to model time features, to capture the correlations within both frequency and time domains. Moreover, an attention mechanism with a frequency attention network (F-Attention) and a time attention network (T-Attention) is also embedded into TFNN to focus on the specific frequency bands and time frames highly corresponding to emotions.

In summary, the contributions of our article mainly include the following three points.

1) We propose a TFNN to jointly model the correlations both in the time and frequency domains, capturing more emotion information in the time–frequency representation of the speech signal.
2) We propose a time–frequency attention mechanism, measuring the contribution degree of frequency bands and time frames associated with different emotions, to capture the critical frequency bands and time frames.
3) By visualizing the attentions in the frequency and time domains, we demonstrate the correlations both within frequency bands and time frames under different emotions on three public speech emotion databases, i.e., IEMOCAP, ABC, and CASIA.

The rest of this article is organized as follows. The proposed ATFNN method is described in Section II. Section III describes the experimental settings and analyzes the experimental results. In the end, Section IV concludes this article and discusses some future works.

II. ATFNN FOR SER

Basically, the ATFNN, shown in Fig. 1, is to embed the time–frequency attention strategy (i.e., F-Attention and T-Attention) into the TFNN. Thus, in this section, we first introduce the TFNN, and then describe the ATFNN model in detail.

A. TFNN

The TFNN aims to learn the high(er)-level time–frequency representation from the input spectrogram feature for SER, which includes two modules, i.e., F-Encoder and T-Encoder.

1) F-Encoder: The role of F-Encoder is to characterize the input spectrogram from the frequency domain, including the encoding of frequency information and the correlation modeling within frequency bands. In the Transformer [24], the self-attention is adopted to calculate the correlations between tokens in natural language processing (NLP). Inspired by this point, we utilize self-attention to aggregate the correlations across frequency bands into the frequency-domain encoding. Thus, we propose an F-Encoder based on the revised encoder of the Transformer to learn the frequency-domain representation of emotional speech, shown in Fig. 2, in which the position encoding module is abandoned to ignore the order relationship modeling between frequency bands. The proposed F-Encoder is introduced in detail below.

Given the input log-Mel-spectrogram feature of emotional speech as \( x \in \mathbb{R}^{f \times t \times 1} \), the F-Encoder \( F_c(\cdot) \) transforms the input \( x \) to the frequency-domain encoding \( \hat{x} \in \mathbb{R}^{f \times t \times c} \).
In ATFNN, the log-Mel-spectrogram (F-Encoder) and time-domain encoder (T-Encoder). The time–frequency attention includes frequency attention (F-Attention) and time attention (T-Attention). In ATFNN, the log-Mel-spectrogram $x$ is first encoded as the enhanced frequency feature $x'$ and the enhanced frequency encoding $\hat{x}'$ by F-Attention and F-Encoder (i.e., the encoder of Transformer), respectively. Then, $\hat{x}'$ is transformed to the enhanced time feature $\hat{x}''$ by T-Attention. Finally, $\hat{x}''$ is as the input of T-Encoder (i.e., Bi-LSTM) to generate the enhanced time encoding $h^{'}$ (i.e., time–frequency encoding) for the emotion classification.

The proposed T-Encoder $T_c(\cdot)$ takes the $i$th frame $\hat{x}_i \in \mathbb{R}^{f \times c}$ of the frequency encoding $\hat{x}$ as the input of each time step in Bi-LSTM to learn the time–frequency representation $h$ of emotional speech. In detail, the time-domain encoding process can be defined as $h_i = T_c(\hat{x}_i)$, where $i \in [1, 2, \ldots, t]$ represents the frame index, $\hat{x}_i = \{\hat{x}_i\}_{i=1}^{t}$, and $h_i$ is the $i$th time-step output of the T-Encoder. Notably, since the T-Encoder is based on three-layer Bi-LSTM with 128 hidden nodes in the article, the final time–frequency feature $h \in \mathbb{R}^{1 \times 256}$ is obtained by concatenating the last hidden states in the forward direction $\overrightarrow{h}_t \in \mathbb{R}^{1 \times 128}$ and the last hidden states in reversed direction $\overleftarrow{h}_t \in \mathbb{R}^{1 \times 128}$, which can be represented as $h = [\overrightarrow{h}_t, \overleftarrow{h}_t]$.  

3) TFNN: The TFNN integrates F-Encoder and T-Encoder into the time–frequency joint learning strategy to ensure that the emotion patterns of speech are captured in both frequency and time domains. Specifically, the input log-Mel-spectrogram feature $x$ is first encoded in frequency bands by F-Encoder to obtain the frequency-domain encoding $\hat{x}$, represented as $\hat{x} = F_c(x)$. Then, $\hat{x}$ is used as the input of T-Encoder to capture the long-term dependence between time frames to obtain the time-domain encoding $h$ through a time–frequency joint learning strategy, which can be formalized as $h = T_c(F_c(x))$.

B. ATFNN

In the ATFNN, two attention networks, i.e., F-Attention and T-Attention, are employed in the TFNN to focus on the emotion-related regions in the frequency and time domains.

1) F-Attention: F-Attention aims to capture salient frequency bands contributing to different emotions, which is implemented by a convolution-based block with the kernel size $(1 \times 5 \times c)$, as shown in Fig. 3. To calculate the frequency attention, we first represent the input log-Mel-spectrogram...
where \( x = \{ x_i \}_{i=1}^{t} \) as a set of \( x_i \in \mathbb{R}^{f \times c} \), where \( i \in \{ 1, 2, \ldots, t \} \) is the frame index and \( x_i \) represents the frequency-domain feature of \( i \)-th frame on \( x \). Since the frequency bands related to emotions are within a certain range, e.g., high frequency, intermediate frequency, and low frequency \([18], [21]\), we employ five frequency bands of \( x_i \) as a band group for the convolution to preserve the correlations within frequency bands. After the convolution operations, the frequency attention weights \( a_i^f \in \mathbb{R}^{1 \times (f/5) \times c} \) of different frequency band groups can be generated, expressed as \( a_i^f = F_a(x_i) \). Moreover, as shown in Fig. 3, we utilize the frequency attention \( a_i^f \) to perform a weighted average operation on \( x_i \) with respect to the channel \( c \) to produce the attention-based frequency features. Then, we integrate these frequency features into the origin input features \( x_i \) by the elementwise addition operation to produce the enhanced frequency feature \( x_i' \), which can be represented as

\[
x_i' = x_i \oplus (F_a(x_i) \odot x_i) = x_i \odot \left( \frac{1}{c} \sum_{m=1}^{c} a_i^{f,m} \otimes x_i \right)
\]  

(1)

where \( a_i^{f,m} \in \mathbb{R}^{1 \times (f/5)} \) represents the attention weight of the \( m \)-th channel on the \( x_i \) frequency attention weight \( a_i^f \), and \( m \in \{ 1, 2, \ldots, c \} \). Moreover, \( \oplus \) and \( \odot \) represent the broadcasting elementwise multiplication and elementwise addition, respectively. It is noting that, in \( \odot \) of F-Attention, each five frequency bands of \( x_i \) share the same attention weights for duplication.

After that, we embeds the F-Attention into the F-Encoder, called the attentive frequency neural network (AFNN), to produce the enhanced frequency encoding \( \hat{x}_i' \in \mathbb{R}^{f \times c \times c} \). The AFNN can be defined as \( \hat{x}_i' = F_a(x_i') \), where \( \hat{x}_i' \in \mathbb{R}^{f \times c} \) and the enhanced frequency encoding \( \hat{x}'' \in \mathbb{R}^{f \times c \times c} \) is a set of \( \hat{x}_i' \), written as \( \hat{x}' = \{ \hat{x}_i' \}_{i=1}^{t} \).

2) T-Attention: Similar to F-Attention, the T-Attention is also performed by a convolution block with the kernel size \((1 \times 8 \times c)\) to capture salient frames that contribute to emotions, as shown in Fig. 4. To calculate the time-domain attention, we first divide the input log-Mel-spectrogram \( x \) into multiple \( x_j \) in frequency domain, where \( j \in \{ 1, 2, \ldots, f \} \) represents the \( j \)-th frequency band of the input \( x \) and \( x_j \in \mathbb{R}^{f \times c} \) is the time-domain feature of the \( j \)-th frequency band on \( x \), i.e., \( x = \{ x_j \}_{j=1}^{f} \). In T-Attention, eight time frames of \( x_j \) are employed as a frame group for the convolution to obtain the time attention weight \( a_i^t \in \mathbb{R}^{1 \times (t/8) \times c} \), expressed as \( a_i^t = T_a(x_j) \). Then, as shown in Fig. 4, we also employ the time attention \( a_i^t \) to perform a weighted average operation on \( x_j \) with respect to the channel \( c \) to generate the attention-based time features, and integrate them into the enhanced frequency
encoding \( \hat{x}_j \) by the elementwise addition operation to obtain the enhanced time feature \( \hat{x}_j' \in \mathbb{R}^{T \times c} \), which can be written as

\[
\hat{x}_j' = \hat{x}_j' \oplus (T_a(x_j) \otimes \hat{x}_j')
\]

\[
= \hat{x}_j' \oplus \left( \frac{1}{c} \sum_{n=1}^{c} a_{j,n} \otimes \hat{x}_j' \right)
\]

(2)

where \( a_{j,n} \in \mathbb{R}^{1 \times (t/8)} \) represents the attention weight of the \( n \)-th channel on the time attention weight \( a_j', \ n \in \{1, 2, \ldots, c\} \), and the enhanced time feature \( \hat{x}_j'' \in \mathbb{R}^{T \times l \times c} \) is a set of \( \hat{x}_j'' \), i.e., \( \hat{x}_j'' = (\hat{x}_j'')_{j=1}^J \). In \( \otimes \) of T-Attention, each eight time frames of \( \hat{x}_j \) share the same attention weights for duplication. Since the T-Encoder is encoded according to the time frame, the ATFNN can also be expressed as \( \hat{h}_j' = T_e(\hat{x}_j') \), where \( \hat{h}_j' = (\hat{h}_j')_{j=1}^J \) is the enhanced time encoding.

3) ATFNN: Integrating F-Attention and T-Attention into F-Encoder and T-Encoder, respectively, extend TFNN to attention-based TFNN, i.e., ATFNN. With F-Attention, the ATFNN can focus on the key frequency bands related to emotions to generate the enhanced frequency-domain encoding \( \hat{x}' \), expressed as

\[
\hat{x}' = F_e(x \oplus F_a(x)).
\]

Then, the T-Attention is adopted to capture key time frames related to emotions to obtain a more discriminative time-frequency representation \( \hat{h}' \), which can be formalized as

\[
\hat{h}_j' = T_e(\hat{x}' \oplus T_a(\hat{x}'))
\]

\[
= T_e(F_e(x \oplus F_a(x)) \oplus T_a(F_a(x \oplus F_a(x))))
\]

(4)

Finally, the final time-domain encoding \( \hat{h}' \) can be regarded as an enhanced time-frequency representation to be fed into the classifier for SER. Actually, the ATFNN can be regarded as the TFNN that embeds time-frequency attention into the process of time-frequency joint learning.

III. EXPERIMENTS

In this section, we describe the used speech emotion databases, then analyze and discuss the experimental results of our proposed ATFNN.

A. Experimental Databases

Three public emotional speech databases, i.e., the interactive emotional dyadic motion capture database (IEMOCAP) [25], the airplane behavior corpus (ABC) [26], and China emotional database (CASIA) [27], [28], are adopted in our experiments to prove the effectiveness of the proposed ATFNN method.

IEMOCAP is a multimodal database with video, speech, and text scripts, collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). It is recorded in dyadic sessions where ten actors (five females and five males) perform the improvised or scripted scenarios to elicit several emotional expressions, i.e., angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and others. In the article, the improvised sentences with four emotions (i.e., angry, happy, sad, and neutral) are all used to perform the experiments according to [18], which has 2280 speech samples.

ABC is an audiovisual emotion database with the German language collected for the particular target application of public transport surveillance. Also, it has 430 speech samples generated by eight subjects (four females and four males) in gender balance. They were induced to express one of six emotions, i.e., aggressive, cheerful, intoxicated, nervous, neutral, and tired, by the given scripts.

CASIA is a Chinese speech emotion database released by the Institute of Automation of Chinese Academy of Sciences. It totally consists of 9600 wav files with six emotional states, i.e., angry, fear, happy, neutral, sad, and surprise. Notably, in the article, we adopt 1200 utterances of CASIA released publicly for the experiment section. Four volunteers with two males and two females are required to simulate these six emotions to produce 300 utterances for each emotion. To describe the selected databases clearly, we also list the detailed parameters of three databases in Table I.

B. Experimental Protocol

To evaluate the performance of the proposed method, we adopt leave-one-speaker-out (LOSO) cross-validation protocol for three selected databases according to [6] and [7]. Specifically, as shown in Table I, the IEMOCAP, ABC, and CASIA databases consist of 10, 8, and 4 speakers, respectively. Therefore, in our experiments, the speech samples of one speaker are utilized as the testing data, while the samples of other speakers are used as the training data. Since the IEMOCAP includes six sessions, the leave-one-session-out strategy is widely adopted in IEMOCAP database according to [18], where four sessions (eight speakers) are for training, and one session (two speakers) is for testing. Thus, we also select several state-of-the-art methods based on this protocol for comparison [18], [23], [29], [30], as shown in Table II.

Besides, we also utilize two evaluation metrics [6], [7], i.e., the weighted average recall (WAR) and the unweighted average recall (UAR), to effectively measure the performance.
TABLE II
EXPERIMENTAL RESULTS WITH WAR AND UAR ON THE THREE PUBLIC SPEECH EMOTION DATABASES
(I.E., IEMOCAP, ABC, AND CASIA), WHERE THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

| Database | Experimental Protocol | Comparison Method                      | Accuracy(%) |
|----------|-----------------------|----------------------------------------|-------------|
|          |                       | DNN-HMM_SGMM-Ali. [31]                  | 62.28       | 58.02       |
|          |                       | CNN+LSTM Model [18]*                   | 68.80       | 59.40       |
|          |                       | CNN_GRU-SeqCap [29]*                   | 72.73       | 59.71       |
|          |                       | FCN+Attention Model [23]*              | 70.40       | 63.90       |
|          |                       | Model-3_Fusion [30]*                   | 72.34       | 58.31       |
| IEMOCAP  | Leave One Session/Speaker Out (LOSO) (5 Sessions or 10 Speakers) | ADARL [32] | 73.02       | 65.86       |
|          |                       | ATFNN (ours)                           | 73.81       | 64.48       |
|          |                       | LLDs+HMM/GMM [6]                       | 57.70       | 48.80       |
|          |                       | RCSR-LPDA [33]**                       | N/A         | 49.40       |
|          |                       | LLDs+SVM [6]                           | 61.40       | 55.50       |
|          |                       | GerDA [7]                              | 61.50       | 56.10       |
|          |                       | ATFNN (ours)                           | 68.84       | 57.57       |
| ABC      | Leave One Speaker Out (LOSO) (8 Speakers) | LLD+DR [34]                            | 39.50       | 39.50       |
|          |                       | DNN+ELM [35]                           | 41.17       | 41.17       |
|          |                       | HoWSF [36]                             | 43.50       | 43.50       |
|          |                       | DTPM [9]                               | 45.42       | 45.42       |
|          |                       | ATFNN (ours)                           | 48.75       | 48.75       |
| CASIA    | Leave One Speaker Out (LOSO) (4 Speakers) | LLDs+HMM/GMM [6]                       | 57.70       | 48.80       |
|          |                       | RCSR-LPDA [33]**                       | N/A         | 49.40       |
|          |                       | LLDs+SVM [6]                           | 61.40       | 55.50       |
|          |                       | GerDA [7]                              | 61.50       | 56.10       |
|          |                       | ATFNN (ours)                           | 68.84       | 57.57       |
|          |                       | LLD+DR [34]                            | 39.50       | 39.50       |
|          |                       | DNN+ELM [35]                           | 41.17       | 41.17       |
|          |                       | HoWSF [36]                             | 43.50       | 43.50       |
|          |                       | DTPM [9]                               | 45.42       | 45.42       |
|          |                       | ATFNN (ours)                           | 48.75       | 48.75       |

* denotes that the used experiment protocol is leave-one-session-out, i.e., four sessions (8 speakers) are for training and one session (2 speakers) is for testing, according to [18]; ** denotes that the samples of 4 speakers are for training and the ones of the other 4 speakers are for testing according to [33]; 'N/A' indicates the results are not reported in the corresponding paper.

of the proposed method, which are commonly adopted in SER tasks. WAR is known as the “normal” recognition actuary, while UAR reflects the classwise recognition accuracy defined as the sum of recalls for each class divided by the number of classes. Because the selected IEMOCAP and ABC databases are both class-imbalanced in our experiments, the WAR and UAR can evaluate the comparison methods more comprehensively.

C. Experimental Setting

To perform experiments conveniently, we make the preprocessing for the speech samples. In detail, following the works in [8], [9], and [18], speech utterances are all resampled to 16000 Hz and divided into small segments with 128 frames (i.e., 20 ms), which can not only preserve the completeness of speech emotion but also augment the dataset. Then, the log-Mel-spectrogram is extracted by the short-time Fourier transform (STFT), in which 20-ms Hamming window size with 50% frame overlapping is adopted and 512-point FFT is used on each frame. Besides, the number of Mel-filter bands is set as 80.

The proposed ATFNN is implemented by the deep-learning framework of Pytorch with NVIDIA GeForce RTX3090 GPUs, which is trained from scratch with the batch size of 128 and optimized by the Adam Optimizer with the initialized learning rate of 0.0005.

D. Results and Analysis

1) Results on IEMOCAP: To compare the proposed ATFNN model with other methods, we select several state-of-the-art works on the IEMOCAP database, i.e., DNN-HMM using an alignment strategy generated from the subspace-based Gaussian mixture model-based HMMs (DNN-HMM_SGMM-Ali) [31], three convolution layers and LSTM with 10-Hz grid resolution (CNN + LSTM Model) [18], GRU layer upon CNN layers with sequential capsules (CNN_GRU-SeqCap) [29], attention-based fully convolutional network (FCN + attention model) [23], fusion model of acoustic and linguistic features (Model-3_fusion) [30], and an adaptive domain-aware representation learning-based model (ADARL) [32].

The experimental results on the IEMOCAP database are shown in Table II. From these results, it is obvious that our proposed ATFNN achieves the state-of-the-art performance. Specifically, our ATFNN obtains the best result on WAR (73.81%) than all comparison methods, and it also obtains suboptimal results on UAR (64.48%) than ADARL (65.86%) [32]. It is worth noting that, since ADARL adds a domain adaptation strategy to eliminate the discrepancy between speakers of the training and testing data, it obtains better performance than ATFNN on UAR. Nevertheless, ADARL depends on the hypothesis that the distribution between training and testing data is within a certain upper error bound, thus its universality is worse than our ATFNN.

Besides, the confusion matrix of ATFNN reported in Fig. 5(a) reveals the detailed performance on IEMOCAP, where ATFNN achieved high recognition accuracies on three emotions, i.e., angry, neutral, and sad, while obtains poor performance on happy. It is because IEMOCAP is an extremely class imbalance database shown in Table I, where happy has the smallest number of samples while neutral has the largest...
number of samples. Therefore, this situation may affect the trained classifier to make it easily recognize happy as neutral.

2) Results on ABC: For the comparison purpose, we also select several public works on the ABC database, i.e., LLDs with HMM/GMM (LLDs + HMM/GMM) [6], random anchor points generalized spectral regression with locally penalized discriminant analysis (RGSR-LPDA) [33], LLDs with SVM (LLDs + SVM) [6], and generalized discriminant analysis (GerDA) based on DNNs [7].

The results on the ABC database, as shown in Table II, also demonstrate that our proposed ATFNN is superior both on W AR (68.84%) and U AR (57.57%) to other comparison methods not only on traditional methods (i.e., LLDs + HMM/GMM, RGSR-LPDA, and LLDs + SVM), but also on deep learning-based methods (i.e., GerDA), which proves that the ATFNN can extract more discriminative and instance-adaptive representation of speech emotion on the ABC database. Moreover, we also report the confusion matrix on the ABC database, shown in Fig. 5(b). It is clear that the ATFNN have high recognition accuracies on four emotions, i.e., aggressive, cheerful, nervous, and neutral, while have poor performance on two emotions, i.e., intoxicate and tired. Specifically, on the ABC, the recognition of intoxicate is easily confusing with cheerful and neutral, whereas tired is more recognized as nervous and neutral. The cause of these results on ABC is similar to reasons for IEMOCAP, that is, the numbers of speech samples on intoxicate and tired are obviously smaller than other emotions, leading to influencing the trained classifier. In addition, another reason may be that the arousal and valance of intoxicate and cheerful are close enough, resulting in the confusing recognition.

3) Results on CASIA: For the fair comparison on CASIA, four state-of-the-art methods are chosen in the experiments, including LLDs with a dimension reduction combining PCA and LDA (LLD + DR) [34], DNNs with the extreme learning machine (DNN_ELM) [35], weighted spectral feature learning based on local Hu moment (HuWSF) [36], and deep convolutional neural network with discriminant temporal pyramid matching (DTPM) [9].

The experimental results on the CASIA are reported in Table II, which reveal that our proposed ATFNN achieves the highest recognition accuracies than other comparison methods in terms of both W AR (48.75%) and U AR (48.75%). In detail, deep learning-based methods, i.e., DNN_ELM and DTPM, outperform the LLDs-based traditional method (e.g., LLD + DR), while our ATFNN is significantly superior than the current state-of-the-art methods. Note that since the number of samples used in each emotion of the CASIA in this article is equal, the W AR results are equal to U AR ones. Furthermore, the confusion matrix of ATFNN on the CASIA is also calculated to verify the detailed performance of our proposed method, which is given in Fig. 5(c). From these results, we can observe that our ATFNN achieves the high performance on four emotions, i.e., angry, happy, neutral, and sad, whereas has a certain confusion in the recognition of other two emotions, i.e., fear and surprise. Specifically, fear is easier to be confused with sad, while surprise is always confused with angry and happy. The reason may be that fear and sad are relatively close in valence and arousal, causing two emotions to be induced by each other sometimes, while surprise, angry, and happy are the high arousal emotions, which may affect the recognition of each other slightly.

4) Discussions of Experimental Results: Throughout all experiments on IEMOCAP, ABC, and CASIA, some results are also worthy of discussing, from which we can investigate the applicability and limitation of our proposed ATFNN. First, two selected databases, i.e., IEMOCAP and ABC, are class-imbalanced, resulting in the discrepancy between W AR and U AR. Moreover, the recognition of different emotions has obvious confusion, as shown in Fig. 5, where 58.27% of speech samples on happy are recognized as neutral in IEMOCAP, and 45.45% of samples on intoxicate in ABC are recognized as cheerful. These issues may be caused by the following reasons: one reason is that these emotions are relatively close in valence or arousal, and the other reason is that the long-tailed distribution of speech samples leads the classifier being biased toward emotions with more samples. Next, although the speech samples of CASIA are
The proposed framework, the additional experiments are implemented for different architectures of ATFNN, and the results in terms of W AR and U AR are reported in Table III, where AFNN and TFNN represent the FNN with F-Attention and the TNN with T-Attention, respectively, described in Section II. From the ablation results, obviously, we can observe that the attention-based time–frequency joint learning method (i.e., ATFNN) has a better performance than the time–frequency joint learning method (i.e., TFNN), which indicates that the attention-based method can focus on the emotion-related parts in the time–frequency domain. In addition, we also observe that the time–frequency-based method (i.e., AFNN) also perform better than the time or frequency-based method (i.e., TFNN), which also demonstrates that time–frequency joint learning can indeed obtain the discriminative representation of speech emotion. It is also interesting to see that, although the frequency-domain representation is meaningful for emotions, the experimental results also show that ATNN performs better than AFNN.

### E. Visualizations of F-Attention and T-Attention

To further investigate the salient regions associated with emotion in the time and frequency domains, we visualize the F-Attention and the T-Attention on the log-Mel-spectrogram, as shown in Fig. 6. For this purpose, high-arousal and positive-valence emotions, e.g., happy, cheerful, and low-arousal and negative-valence emotions, e.g., sad, tired, are selected as comparison for attention visualization [37], [38]. Among them, we choose happy and sad for a comparison on IEMOCAP and CASIA, whereas cheerful and tired are chosen as the alternatives on ABC, which have close valence and arousal with happy and sad, respectively. Fig. 6(a)–(c) shows the attentions under happy and sad on IEMOCAP, where Fig. 6(a)–(c), respectively, corresponds to the visualizations of log-Mel-spectrogram, F-Attention, and T-Attention under happy, while Fig. 6(d)–(f) is the visualizations of above three items under sad. Similarly, the details of Fig. 6(g)–(r) are given in Fig. 6. Note that, as shown on Fig. 6, the black dotted line “--” on each figure represents the Mel-scaled frequency of 50, corresponding to the actual frequency band around 2700 Hz. Therefore, the upper part of the black dotted line means the high-frequency energy activation region, while the lower part represents the middle and low-frequency energy activation regions. In addition, the attention visualizations reveal the specific activations under different emotions, where the colors of visualizations from red to blue represent corresponding activation intensity from large to small.

From the comparison of Fig. 6(a) and (b), it is obvious to observe that there are frequent energy activations of happy above 50 at the Mel-scaled frequency (corresponding to the frequency band around 2700 Hz), which demonstrates that happy has an increased activations in high-frequency energy. Whereas, as shown in Fig. 6(d) and (e), we can clearly find that energy activations of sad are infrequent above 50 at the Mel-scaled frequency, which indicates sad has a decreased activations in high-frequency energy. Similar results also appear in ABC and CASIA, as shown in Fig. 6(g) and (j), Fig. 6(m) and (p), respectively. These visualization results are consistent with the concepts reported in [20], [21], [22]. Note that the results in Fig. 6(j) and (k) also reveal that the F-Attention of tired on ABC has frequent energy activations on high-frequency bands from 10th frame to 40th frame and 125th frame to 250th frame. This is because the used speech sample has obvious sound of “yawning” in these two frame intervals. On the contrary, the sample from 50th frame to 80th frame contains a conventional emotion expression of tired, thus its F-Attention reveals infrequent energy activations above 50 Mel-scaled frequency, which is accorded with the above demonstration. Moreover, the visualization of F-Attention confirms that our proposed method can focus on the frequency-band activations related to emotions, and then combine the

| Network | Ablation Architecture | IEMOCAP (%) | ABC (%) | CASIA (%) |
|---------|-----------------------|-------------|---------|-----------|
|         | Attention | F-Encoder | T-Encoder | WAR | UAR | WAR | UAR | WAR | UAR |
| AFNN    | ✓         | ✓         | ✗        | 70.71 | 59.10 | 58.14 | 48.22 | 32.50 | 32.50 |
| ATNN    | ✓         | ✗         | ✓        | 71.76 | 61.90 | 66.05 | 55.01 | 46.42 | 46.42 |
| TFNN    | ✗         | ✓         | ✓        | 72.99 | 63.60 | 67.21 | 55.99 | 46.67 | 46.67 |
| ATFNN   | ✓         | ✓         | ✓        | 73.81 | 64.48 | 68.84 | 57.57 | 48.75 | 48.75 |

**TABLE III**

Ablation Study of Different Architectures for ATFNN on the Three Public Speech Emotion Databases (i.e., IEMOCAP, ABC, and CASIA), Where the Best Results Are Highlighted in Bold. Moreover, “✓” or “✗” Represents the Network With or Without the Corresponding Architecture.
correlations between these frequency bands into the frequency-domain representation.

As for T-Attention, Fig. 6(c) indicates that happy on IEMOCAP has high activations in 50th frame to 90th frame and 110th frame to 270th frame. These frames correspond to the regions with rich speech contents shown in Fig. 6(a), which are all key regions in frames that highly contribute to speech emotion representation. From the visualization of T-Attention on other emotions or databases, i.e., Fig. 6(f), (i), (l), (o), and (r), it is clear to discover similar results with Fig. 6(c). Also, our proposed T-Attention has few activations on all silent frames, e.g., 0th frame to 150th frame and 360th frame to 384th frame in Fig. 6(d), 220th frame to 256th frame in Fig. 6(g), 80th frame to 120th frame and 210th frame to 480th frame in Fig. 6(j), 70th frame to 120th frame in Fig. 6(m), 40th frame to 60th frame in Fig. 6(p). These salient results further emphasize that our proposed T-Attention effectively focuses on the key regions related to emotions in time frames.
IV. CONCLUSION

In this article, we propose an ATFNN to extract the discriminative representations of speech emotions for the SER for the SER. We first design a TFNN, integrating a frequency-domain encoder and a time-domain encoder into the time–frequency joint learning strategy, to model the correlations within frequency bands and across time frames. Then, a frequency attention network and a time attention network are also embedded into the TFNN to capture the specific frequency bands and time frames related to emotions. Experimental results on the three public databases, i.e., IEMOCAP, ABC, and CASIA, prove that our ATFNN outperforms the state-of-the-art methods. Moreover, the visualization of attentions reveals that the ATFNN can capture the regions related to emotions on the frequency and time domains. Furthermore, the precise correspondence between different emotions and frequency bands should be deeply explored in future research to improve the robustness and discriminability of speech emotion features.

REFERENCES

[1] B. Schuller and A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech And Language Processing, Hoboken, NJ, USA: Wiley, 2013.
[2] C. Busso, M. Bulut, S. Narayanan, J. Gratch, and S. Marsella, “Toward effective automatic recognition systems of emotion in speech,” Social Emotions in Nature and Artifact: Emotions in Human and Human–Computer Interaction, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford Univ. Press, 2013, pp. 110–127.
[3] J. Guo, L. Wang, J. Dang, E. Chng, and S. Nakagawa, “Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition,” Speech Commun., vol. 136, pp. 118–127, Jan. 2022.
[4] M. Li et al., “Contrastive unsupervised learning for speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 6329–6333.
[5] B. Schuller, B. Vlasonko, F. Eyben, G. Rigoll, and A. Wendemuth, “Affective emotion recognition: A benchmark comparison of performances,” in Proc. IEEE Workshop Autom. Speech Recognit. Understand., Dec. 2009, pp. 552–557.
[6] A. Gawel, C. Mathez, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5688–5691.
[7] S. Zhang, S. Zhang, T. Huang, and W. Guo, “Speech emotion recognition using deep convolutional neural networks,” IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[8] X. Xu, X. Xu, X. Xing, and D. Huang, “Adaptive domain-aware representation learning for speech emotion recognition,” in Proc. Interspeech, Oct. 2020, pp. 4089–4093.
[9] Z. Liu, Q. Xue, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, “Speech emotion recognition based on an improved brain emotion learning model,” Neurocomputing, vol. 309, pp. 145–156, Oct. 2018.
[10] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The Munich versatile and fast open-source audio feature extractor,” in Proc. 18th ACM Int. Conf. Multimedia (MM), 2010, pp. 1459–1462.
[11] F. Eyben, M. Wöllmer, and B. Schuller, “Speech emotion recognition with dual-sequence LSTM architecture,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6474–6478.
[12] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Commun., vol. 116, pp. 56–76, Jan. 2020.
[13] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech emotion classification using attention-based LSTM,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 27, no. 11, pp. 1675–1685, Jul. 2019.
[14] J. Li, A. Mohamed, G. Zweig, and Y. Gong, “LSTM time and frequency recurrence for automatic speech recognition,” in Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU), Dec. 2015, pp. 187–191.
[15] L. Guo, L. Wang, J. Dang, E. S. Chng, and S. Nakagawa, “Learning affective representations based on magnitude and dynamic relative phase for speech emotion recognition,” in Proc. IEEE/ACM Trans. Audio, Speech, Language Process., vol. 18, no. 5, pp. 1089–1093.
[16] J. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[17] S. Wu, T.-H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, 2011.
[18] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1–11.
[19] J.-C. Wang, Y.-H. Chin, B.-W. Chen, C.-H. Lin, and C.-H. Wu, “Speech emotion verification using emotion variance modeling and discriminant scale-frequency maps,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 10, pp. 1552–1562, Oct. 2015.
[20] J. T. F. L. M. Zhang and H. Jia, “Design of speech corpus for Mandarin text-to-speech,” in Proc. Blizzard Challenge Workshop, 2008, pp. 1–4.
[21] X. Wu et al., “Speech emotion recognition using capsule networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6695–6699.
[22] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, “Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2020, pp. 1–5.
[23] X. Xu, X. Xu, X. Xing, and D. Huang, “Adaptive domain-aware representation learning for speech emotion recognition,” in Proc. Interspeech, Oct. 2020, pp. 4089–4093.