IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 16, 2019, accepted January 11, 2020, date of publication January 20, 2020, date of current version January 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967791

Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends

ZHICHAO PENG[1,2], XINGFENG LI[1], ZHI ZHU[1], MASASHI UNOKI[1], JIANWU DANG[1,2,3], AND MASATO AKAGI[1]

[1]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan
[2]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[3]Peng Cheng Laboratory, Shenzhen 518055, China

Corresponding author: Zhichao Peng (zcpeng@jaist.ac.jp)

This work was supported in part by the JSPS KAKENHI Grant under Grant 16K00297, and in part by the Research Foundation of Education Bureau of Hunan Province, China, under Grant 18A414.

ABSTRACT Emotion information from speech can effectively help robots understand speaker's intentions in natural human-robot interaction. The human auditory system can easily track temporal dynamics of emotion by perceiving the intensity and fundamental frequency of speech, and focus on the salient emotion regions. Therefore, speech emotion recognition combined with the auditory mechanism and attention mechanism may be an effective way. Some previous studies used auditory-based static features to identify emotion while ignoring the emotion dynamics. Some other studies used attention models to capture the salient regions of emotion while ignoring cognitive continuity. To fully utilize the auditory and attention mechanism, we first investigate temporal modulation cues from auditory front-ends and then propose a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks (ASRNNs) for emotion recognition. Our experiments on the IEMOCAP and MSP-IMPROV datasets indicate that the proposed method can be effectively used to recognize the emotions of speech from temporal modulation cues. The subjective evaluation shows that the attention patterns of the attention model are basically consistent with human behaviors in recognizing the emotions.

INDEX TERMS Auditory front-ends, 3D convolutions, joint spectral-temporal representations, attention-based sliding recurrent networks, speech emotion recognition.

I. INTRODUCTION

Speech is the most natural way for communication between humans and robots. The key point of effective communication is to make robots or virtual agents understand speakers' true intentions. However, only using the linguistic information is by no means sufficient enough for understanding of intentions. The vocal emotion information as a kind of non-linguistic information can significantly help robots or virtual agents to understand speakers' true intentions. Therefore, speech emotion recognition (SER) is the research hotspot in natural human-robot interaction (HRI). Nevertheless, effective SER is still a very challenging problem, partly due to the cultural differences, various expression types, context, ambient noise, etc.

In most of the past SER, low-level descriptors (LLDs) were extracted from speech and were used to classify different emotion states by means of the conventional machine learning methods such as hidden Markov models (HMM), Gaussian mixture model (GMM), and support vector machine (SVM) [1]. However, it is still difficult to find the salient feature set from LLDs to recognize distinct emotions, because of the aforementioned challenging factors. The human auditory system can easily perceive the intensity and fundamental frequency of speech, and can track temporal dynamics of emotion from the perceived information and focus on the salient emotion regions. Therefore, speech emotion recognition combined with the auditory mechanism of auditory

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

front-ends and attention mechanism of auditory back-ends may be an effective way.

In auditory front-ends, temporal modulation cues are obtained using auditory filtering of speech signal and modulation filtering of temporal amplitude envelope. These cues contain rich spectral-temporal information to perceive the variations of intensity, duration and pitch of speech [2] and have been widely used in sound-texture perception [3], speaker-individuality perception [4], speech recognition [5], and emotion recognition [6], [7]. Most studies extracted the modulation spectral features (MSFs) from temporal modulation cues by calculating the spectral centroid, flatness, skewness, kurtosis, and other statistical features. Wu *et al.* [6] showed that the MSFs perform better than the traditional acoustic features such as Mel frequency cepstral coefficient (MFCC) and perceptual linear predictive (PLP) coefficient for SER. Zhu *et al.* [7], [8] further confirmed that the MSFs contribute to the perception of vocal emotion. However, the MSFs are only calculated in each modulation channel and produce time-averaged static features in those studies. Since emotion in speech is often communicated by varying temporal dynamics in the signal, the temporal dynamics are very important factors in emotion recognition. The MSFs cannot reflect the real emotion in speech since it lost the important temporal cues. For these reasons, we should extract the joint spectral-temporal features from temporal modulation cues to accurately describe emotion dynamics.

Recent convolutional neural networks (CNNs) show powerful abilities of feature learning and have been used for acoustic modeling and feature extraction for SER. As human auditory system responds to joint spectral-temporal patterns in the speech signal rather than temporal-only or spectral-only patterns [9]. Inspired by auditory signal processing, in our previous study [10], we proposed an end-to-end SER system using 3D CNNs to learn a joint spectral-temporal feature from temporal modulation cues containing acoustic frequency components, modulation frequency components, and temporal features. The modulation frequency components consist of six filters spaced on a logarithm scale from 2 to 64 Hz. Such modulation frequency components include the local information about variations of intensity and duration. However, it did not take into account of obtaining the periodicity information about F0 from the modulation frequency band. The frequency band between about 50 and 500Hz is related to the periodicity information about F0, which has been shown to be important for speech perception [11]. To obtain both the local features and periodicity information, in this study, we improve the 3D convolution model by increasing the modulation filters and reducing the convolutional kernel size.

To capture the variations of local features and periodicity information from the feature sequence, we need to extract utterance-level features for classifying emotional speeches through time series modeling. Long short-term memory recurrent neural networks (LSTM-RNNs) have powerful abilities of time series modeling to handle temporal dynamic information. LSTM can effectively capture the long-range time dependencies for sequence classification. However, it cannot avoid the slow training speed caused by backpropagation-through-time (BPTT) in long sequences. To reduce the training cost, in [10], the time sequence is divided into non-overlapping subsequences in extraction of segment-level features. These discontinuous segment-level features cannot fully reflect the dynamic changes of real emotions. From a cognitive point of view, people can obtain important information by scanning the temporal sequence continuously and transmit it for higher-level processing. In addition, people have superior abilities in paying attention to the emotional regions meanwhile ignoring the emotionless regions. Most of studies did not take into account of the human mechanism how to focus on the emotional segments while ignoring the emotionless segments. An utterance consists of a number of voiced and unvoiced segments. The voiced segments can express emotion more than the unvoiced ones. It is unknown what kind of acoustic features attract human to pay more attention on the salient emotional regions. Therefore, we will investigate the relation of the acoustic features and human attention mechanism, and propose a sliding recurrent method to realize the attention mechanism. In the temporal attention method, the continuous segment-level internal representations are extracted by a sliding window, and are used to capture the salient emotional regions.

To fully utilize the human auditory mechanism and attention mechanism, in this study, we begin with the investigation of temporal modulation cues from auditory front-ends and then find out a method to capture the salient emotional regions. Based on the achievements, we propose a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks (ASRNNs) as the back-ends of the SER system. To show the benefit of the proposed model, we evaluate it on the IEMOCAP [12] and MSP-IMPROV [13] datasets by comparing various models with the proposed model. Our results show that the proposed model can achieve better results compared with traditional model on both datasets. We also conduct the subjective evaluation to investigate the relevance between the attention patterns of the temporal attention model and human attention in perceiving emotional speech.

The main contributions of this work are summarized as follows:

1) Inspired by the auditory signal processing and temporal attention mechanism, we propose a speech emotion recognition system that combines auditory perception-based front-end and attention-based back-end. In this system, the front-end is used to generate temporal modulation cues, and the attention-based back-end is used to identify the emotional states in natural speech.

2) We propose a 3D convolution model to obtain both the local features and periodicity information of emotional speech by a joint spectral-temporal feature learning from the temporal modulation cues.

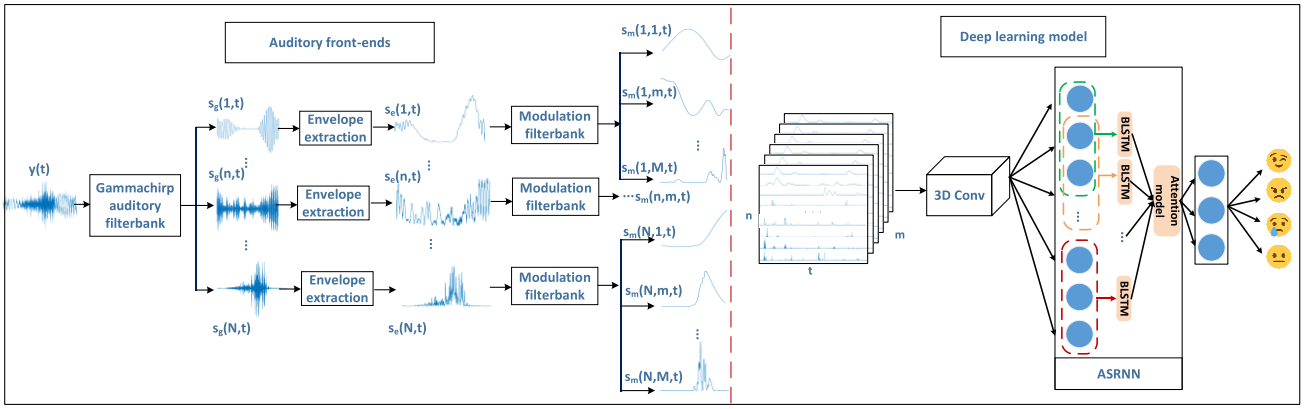

FIGURE 1. Speech emotion recognition system with auditory front-ends.

3) We propose an ASRNN to continuously scan the temporal sequence and focus on the emotional region. In this neural network, the continuous segment-level internal representations are extracted by a sliding window and focus on the salient emotion regions using a temporal attention model.

The rest of the paper is organized as follows. In Section 2, we introduce the auditory front-ends to produce temporal modulation cues. Section 3 details the 3D convolutions to learn a joint spectral-temporal feature representation from those cues and ASRNNs to focus on the salient emotion regions. In Section 4, we also investigate the impacts of experiments on different situations. We discuss the implications of this study in Section 5. Finally, we draw conclusions in Section 6.

II. PROPOSED AUDITORY FRONT-ENDS OF EMOTION RECOGNITION SYSTEM

A. OVERVIEW OF EMOTION RECOGNITION SYSTEM

An overview of the proposed SER system is illustrated in Fig. 1. The auditory front-ends of this system are used to functionally simulate the signal processing in the auditory system from the cochlea through the thalamus, as depicted in the left part of Fig. 1.

The auditory front-ends are composed of three parts: auditory filterbank, temporal envelope extraction and modulation filterbank. The auditory filterbank is responsible to decompose speech signals into acoustic frequency components as a function of the acoustic frequency analyzer in the cochlea. In this study, we use Gammachirp filterbank [14] as the auditory filterbank because this filter is adequate for reproducing psychophysically estimated human auditory filters over a wide range of center frequencies and levels [15], [16]. Furthermore, temporal envelope extraction from the acoustic frequency components is used to effectively simulate the mechanical-to-neural signal transduction in the inner hair cells (IHCs).

Modern psychophysical models of temporal modulation processing suggest that the temporal envelope is processed by joint spectral-temporal modulations [17]. The spectral-temporal modulation contains the 3D modulated spectrum with dynamic peaks, which relates directly to speech perception [9]. Hence, the modulation filterbank is introduced to generate 3D spectral-temporal representations from the temporal envelope.

The back-ends of this system are depicted in the right part of Fig. 1. 3D convolutions are firstly used to extract joint frame-level features including not only variations information of intensity and duration but also the periodicity information. Further, ASRNNs are used to focus on the salient emotional regions by extracting segment-level features in a sliding window manner and utterance-level features with a temporal attention model.

B. FRONT-END SIGNAL PROCESSING

In the auditory front-end, the emotional speech signal $y(t)$ is first filtered by a bank of Gammachirp auditory filters. The output of the nth channel signal is given by

$$s_{\mathrm{g}}(n,t) = g_{\mathrm{c}}(n,t) * y(t), \quad 1 \le n \le N, \tag{1}$$

where $g_{\mathrm{c}}(n,t)$ is the impulse response of the nth channel, t is the sample number in the time domain, N is the number of channels in the auditory filterbank, and $*$ denotes the convolution. The center frequencies of these filters are proportional to their bandwidths, which in turn are characterized by the equivalent rectangular bandwidth ($\mathrm{ERB_N}$) [18]:

$$ERB_N(f_{\mathrm{n}}) = \frac{f_{\mathrm{n}}}{Q_{ear}} + B_{min}, \tag{2}$$

where f_{n} is the center frequency of the nth filter, Q_{ear} is an asymptotic filter quality at large frequencies, B_{min} is minimum bandwidth at low frequencies. Filter quality is a measure of its center frequency divided by the bandwidth. The most widely accepted is provided by [19] in which Q_{ear} and B_{min} are 9.26449 and 24.7, respectively. This impulse response of Gammachirp filter is the product of the Gamma distribution and sinusoidal tone.

$$g_{\mathrm{c}}(n,t) = \mathrm{A}t^{a_1-1}\exp\left(-2\pi w_f ERB_N(f_{\mathrm{n}})\,t\right) \times \cos\left(2\pi f_{\mathrm{n}} t + c_1 \ln(t) + \varphi\right), \tag{3}$$

where $At^{a_1-1}\exp(-2\pi w_f ERB_N(f_n)t)$ is the amplitude term represented by the Gamma distribution, A, a_1 and w_f are the amplitude, filter order, and bandwidth of the filter, respectively. The $c_1 \ln(t)$ term is the monotonic frequency modulation term, φ is the original phase, and $ERB_N(f_n)$ is a bandwidth of the auditory filter in f_n. The chirping properties of the Gammachirp filter are largely determined by those of its "passive" asymmetric filter at all levels and have been shown to fit those of auditory nerve fibers well [14].

The envelope is extracted using the Hilbert transform to calculate the instantaneous amplitude $s_e(n,t)$ of the nth channel signal. The $s_e(n,t)$ is computed from $s_g(n,t)$ as the magnitude of the complex analytic signal $\widehat{s_g}(n,t) = s_g(n,t) + j\mathcal{H}\{s_g(n,t)\}$, where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. Hence,

$$s_e(n,t) = \left|\widehat{s_g}(n,t)\right| = \sqrt{s_g^2(n,t) + \mathcal{H}^2\left\{s_g(n,t)\right\}}. \quad (4)$$

Furthermore, the mth modulation filter in the nth channel signal is used to obtain the spectral-temporal modulation signal $s_m(n,m,t)$.

$$s_m(n,m,t) = m_f(m,t) * s_e(n,t), \quad 1 \le m \le M, \quad (5)$$

where $m_f(m,t)$ is the impulse response of the modulation filterbank and M is the number of channels in the modulation filterbank.

This type of signal generates a frequency-domain-specific time-domain signal for each sub-channel and many sub-channels comprise the 3D spectral-temporal representation. Due to the high time-resolution of the spectral-temporal representations, a reduction in the number of samples for the time domain has to be carried out. The reduction in the time-resolution is simply carried out by downsampling spectral-temporal representations with an 800-Hz rate. This operation reduces the sequence length by a factor of 20.

C. MODULATION SPECTRAL REPRESENTATIONS

Figure 2 shows the different emotion examples of the modulation spectral representation with 32 acoustic channels and nine modulation channels from the IEMOCAP dataset. Each utterance comes from the same speaker, named Ses01F_impro05_F009 (Angry), Ses01F_impro03_F001 (Happiness), Ses01F_impro04_F000 (Neutral emotion), and Ses01F_impro02_F005 (Sadness), respectively. The y-axis and x-axis of these representations are acoustic and modulation channels, respectively. Both channels are spaced on a logarithm-scale frequency. Modulated signals with standard deviation are projected into the modulation and acoustic frequency space. Panels (a) to (d) in Fig. 2 show the modulation spectral representations of anger, happiness, neutral emotion and sadness, respectively. As slow modulation frequency, particularly below 16 Hz (modulation channel equals to 4), can extract local information about variations of intensity, duration, attack, decay, and segmental cues of speech [20]. From these panels, we can find that the different emotion has different low frequency modulation information, suggesting they could be discriminated from each other. In [10], we therefore used six modulation filters to extract low frequency information (below 64 Hz) for emotion recognition.

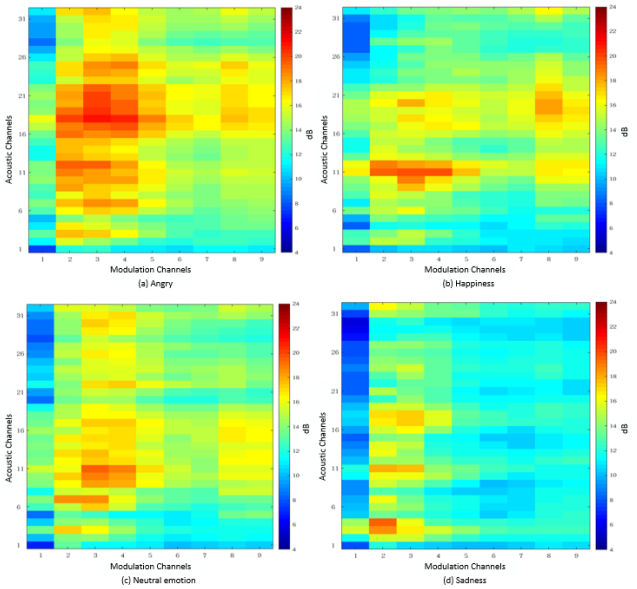

FIGURE 2. Different emotion examples of the modulation spectral representation with 32 acoustic channels and nine modulation channels from the IEMOCAP dataset.

Although fast modulation frequency is less important than slow modulation frequency, it still contains the periodicity information to reflect emotional changes. Figure 2 also shows that the periodicity information is retained between the seventh and ninth modulation channels. In addition, for the same fast modulation frequency, it shows that the acoustic frequency of anger and happiness is higher than that of sadness and neutral emotion. For this reason, we use nine modulation filters with upper limit of modulation frequency (512 Hz) instead of six filters to obtain periodicity information for emotion recognition.

III. METHODS

As illustrated in the right a of Fig. 1, the proposed backends of the SER system are composed of two components: 3D convolutional model and attention-based sliding recurrent networks.

A. 3D CONVOLUTIONAL MODEL

Since deep convolutional model keeps the spectral-temporal translation invariance for speech signal processing, it is often used to extract high-level features for speech emotion recognition. Most studies used CNNs to extract 2D feature representations from speech spectrograms [21], [22] or Mel-scaled filterbank representation [23], [24]. Recently some studies proposed 3D convolution models to better capture the spectral-temporal relationship of the feature representations for emotion recognition. Chen *et al.* [25] proposed attention-based CRNN from a 3D feature representation by computing the log Mel-spectrogram with deltas and delta-deltas for emotion recognition. Kim et al. proposed deep

TABLE 1. 3D convolutional neural networks architecture.

Layer	Input size	Output size	Kernel	Stride
Conv1	32x9x6000	32x9x3000	2x2x4	1x1x2
Pool1	32x9x3000	16x4x3000	2x2x1	2x2x1
Conv2	16x4x3000	16x4x3000	2x2x4	1x1x1
Pool2	16x4x3000	8x2x1500	2x2x2	2x2x2
Conv3	8x2x1500	8x2x1500	2x2x4	1x1x1
Pool3	8x2x1500	4x2x750	2x1x2	2x1x2
Reshape	4x2x750	750x512		

3D CNNs for spectral-temporal feature learning by dividing the speech signal into several sub-segments and these sub-segments contain 2D feature maps with 256 points log-spectrogram for every 20 ms [26]. In this study, the temporal modulation cues from the auditory front-ends contain 3D spectral-temporal representation. The back-ends of the SER system are responsible for extracting high-level features from the 3D representation. CNNs have superior feature extraction power inspired from biological neural networks and can extract high-level local feature representations using the spectral-temporal receptive field of the neuron. Therefore, we use 3D CNNs to learn a joint spectral-temporal feature from the 3D representation for obtaining local features and periodicity information.

The architecture of 3D CNNs is described in Table 1. The first convolutional layer (Conv1) is used to extract 3D features that are composed of acoustic frequency, modulation frequency, and time sequences. These features are fed into the next two convolutional layers (Conv2 and Conv3) to model high-level feature representations for time series. The data format of the input and output data is designed as ''D × H × W'', where D, H, and W are the data in the acoustic channels (depth), modulation channels (height), and time sequence (width), respectively. In this study, the input size is set as $32 \times 9 \times 6000$ and the size of the kernels is $2 \times 2 \times 4$. To reduce computational complexity, the stride for Conv1 is set to $1 \times 1 \times 2$, and that for the other convolutional layers is set to $1 \times 1 \times 1$. Each convolutional layer includes batch normalization and rectified linear unit (ReLU) operations. Batch normalization is used to accelerate training of deep network [27]. The first pooling layer (Pool1) before conv2 has a kernel size of $2 \times 2 \times 1$ and stride of $2 \times 2 \times 1$ with max-pooling operation. The second pooling layer (Pool2) has a kernel size of $2 \times 2 \times 2$ and stride of $2 \times 2 \times 2$. This means that spectral-temporal pooling is executed on Pool2. The third pooling layer (Pool3) has a kernel size of $2 \times 1 \times 2$ and stride $2 \times 1 \times 2$. This means that the acoustic frequency channel and temporal pooling is executed while the modulation frequency channel remains on Pool3. The max-pooling operations in each pooling layer is used to extract robust features against background noise, especially for the waveform signals. These three pooling layers reduce the output size of the time sequence by a factor of 20 on the temporal length. This means that the 3D convolution only learns the frame-level features in 22.5ms for each point. The feature maps of the three convolution layers are 20, 32, and 64, respectively. Finally, we obtain the output of Pool3 with the shape of $750 \times 4 \times 2 \times 64$ after transposing the axis of the tensor then reshape it to 2D shapes of 750×512.

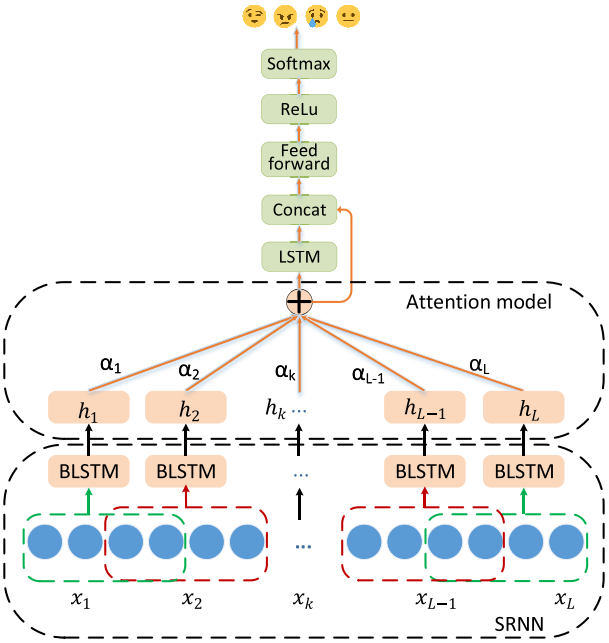

FIGURE 3. Attention-based sliding recurrent networks.

B. ATTENTION-BASED SLIDING RECURRENT NEURAL NETWORKS

Part of the attention system of the brain is involved in the control of thoughts, emotions, and behavior. In human auditory system, selective auditory attention tracks the temporal dynamics of emotion by continuous scanning and encoding of the speech signals [28]. Inspired by the selective auditory attention in auditory system, we propose an ASRNN model to seize the emotional parts from temporal dynamics information in speech. Among them, a sliding window is used to extract the continuous segment-level emotional features containing temporal dynamics information. Then, a temporal attention model is used to capture the important information related to emotion in each utterance.

1) SLIDING RECURRENT NEURAL NETWORKS

The sliding recurrent neural networks (SRNNs) are used to continuously extract the intermediate segment-level representations for short-term sequence depicted in Fig. 3. The input of the SRNNs is T × D, where T represents the total length of the time sequence and D represents the feature vector size. x_k is the input to the LSTM block of kth sliding input sequence with Z time frames.

$$x_k = \{x_{(k,1)}, \ldots, x_{(k,Z)}\}, \quad x_{(k,t)} \in \mathbb{R}^D, \ 1 \le t \le Z \quad (6)$$

Each x_k is fed frame-by-frame into the LSTM units. The formulation of LSTM with peephole connections can be described by the following equations:

$$i_{(k,t)} = \sigma(W_{ix}x_{(k,t)} + W_{ih}h_{(k,t-1)} + W_{ic}c_{(k,t-1)} + b_i) \quad (7)$$
$$f_{(k,t)} = \sigma(W_{fx}x_{(k,t)} + W_{fh}h_{(k,t-1)} + W_{fc}c_{(k,t-1)} + b_f) \quad (8)$$
$$\widetilde{c_{(k,t)}} = tanh(W_{cx}x_{(k,t)} + W_{ch}h_{(k,t-1)} + b_c) \quad (9)$$
$$c_{(k,t)} = f_{(k,t)} \odot c_{(k,t-1)} + i_{(k,t)} \odot \widetilde{c_{(k,t)}} \quad (10)$$
$$o_{(k,t)} = \sigma(W_{ox}x_{(k,t)} + W_{oh}h_{(k,t-1)} + W_{oc}c_{(k,t)} + b_o) \quad (11)$$
$$h_{(k,t)} = o_{(k,t)} \odot tanh\left(c_{(k,t)}\right), \quad (12)$$

where $i_{(k,t)}, f_{(k,t)}, o_{(k,t)}, c_{(k,t)}$, and $h_{(k,t)}$ are the input gate, forget gate, output gate, cell state, and output of the LSTM block, respectively, at the current time step t. The weight matrices W_{i*}, W_{f*}, and W_{o*} transform x_k and hidden state $h_{(k,t-1)}$, respectively, to cell update $\widetilde{c_{(k,t)}}$ and three gates $i_{(k,t)}, f_{(k,t)}$, and $o_{(k,t)}$. Finally, b_i, b_f, b_o are the additive biases of the input gate, forget gate, and output gate, respectively. The set of activation functions consists of the logistic sigmoid function $\sigma(\cdot)$, element-wise multiplication $\odot$, and hyperbolic tangent function tanh$(\cdot)$.

Specifically, we use a bidirectional LSTM (BLSTM) network in this study, where the sequence of received signals is once fed in the forward direction into one LSTM cell, and once fed in backwards into another LSTM cell. The forward LSTM reads the time sequence in its original order and generates a hidden state $fh_{(k,t)} = \{fh_{(k,1)}, \ldots, fh_{(k,Z)}\}$ at each time step. Similarly, the backward LSTM reads the time sequence in its reverse order and generates a sequence of hidden states $bh_{(k,t)} = \{bh_{(k,Z)}, \ldots, bh_{(k,1)}\}$. The last state of the forward and backward LSTM cells carry information of the entire source sequence. We concatenate the last state of the forward and backward LSTM cells to produce the h_k of k sequence.

$$h_k = [fh_{(k,Z)}, bh_{(k,1)}] \quad (13)$$

Each hidden state h_k contains information of each sliding window sequence. The hidden states of the recurrent layer along the different frames of the window are used to compute the extracted features. The output of this layer for each sliding window is the cell state vector of the last time frame in each sliding window. After processing in each sliding window, we shift S time frames to compute the next sliding window with the valid padding. The number of sliding window L is calculated as

$$\mathrm{L} = \lceil(\mathrm{T} - \mathrm{Z})/\mathrm{S}\rceil. \quad (14)$$

The BLSTM has 512 hidden units for both directions in each sliding window. Finally, we create a new sequence with the shape of $L \times 1024$ to put into the attention model. The same parameters of the LSTM cell are used in each sliding sequence, then a new context sequence h is produced.

$$h = \{h_1, \ldots, h_L\}, \quad h_k \in \mathbb{R}^{2D}, \ 1 \le k \le L \quad (15)$$

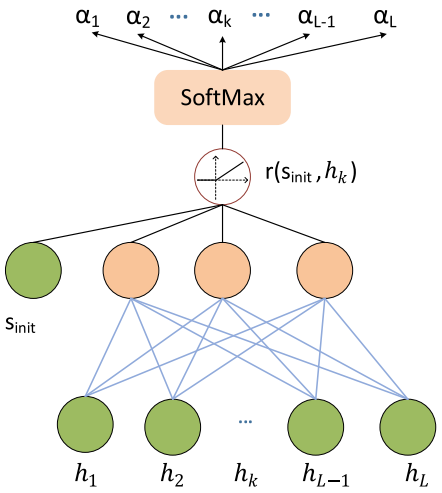

FIGURE 4. Attention weights.

2) TEMPORAL ATTENTION MODEL

Because there are many speech frames that are unrelated to the expressed emotion, such as silence, the attention mechanism is mainly used to focus only on the significant emotional part of the speech signal. Recently, some studies proposed attention models to adjust weights for each of the speech frames depending on their importance based on LLDs using a RNN [29], [30]. The silence regions can be addressed using a voice activity detection (VAD) [31] or by null label alignment [32]. Wang *et al.* [31] proposed an attention model of learning utterance-level representations to improve classification after using a VAD to filter out silence frames and mini-batch training in each utterance. Lee and Tashev [32] extracted high-level representation of emotional states with regard to its temporal dynamics using the BLSTM approach, in which they assume that different frames should have different labels and the label sequence should be alternating between the utterance-level label and a newly introduced NULL state. Neumann and Vu [33] proposed an attentive convolutional neural network (ACNN) to test the emotional discrimination of different feature set. In addition, self-attention based deep model [34], [35] demonstrated the effectiveness to improve the performances for SER. Unlike these studies, we apply a temporal attention model to the sliding window sequence instead of applying one based on LLDs.

Sequence h is fed into feedforward neural networks then concatenated with s_{init}, as depicted in Fig.4. Subsequently, a ReLU is used to produce non-linear transformations $\mathcal{R}(\mathrm{s_{init}}, h_k)$.

$$\mathcal{R}\left(s_{init}, h_k\right) = U_k ReLU\left(s_{init} + W_k h_k + b_k\right), \quad (16)$$

where W_k, U_k are the trainable parameter matrices, b_k is the bias vector, and $\mathrm{s_{init}}$ is the initial hidden state of the sliding recurrent sequence. We use the non-linear function of the ReLU due to its good convergence performance. For each h_k,

the α_k can be computed as follows:

$$\alpha_k = \frac{exp\left(\mathcal{R}(s_{init}, h_k)\right)}{\sum_{l=1}^{L} exp\left(R(s_{init}, h_l)\right)}. \quad (17)$$

We then obtain the attention weights α_k of each sliding sequence from the attention model. The output of the attention layer, *attention_sum*, is the weighted sum of h.

$$attention\_sum = \sum_{k=1}^{L} \alpha_k h_k. \quad (18)$$

The weighted sum of sequence h is fed into a unidirectional LSTM cell to obtain a hidden vector h_s. The features concatenated by h and h_s are fed into feedforward neural networks. Subsequently, we use a ReLU as the activation function, which brings the non-linearity into the networks. Finally, we use the softmax to produce the emotion state distribution. To avoid overfitting when training our networks, we use a dropout rate of 0.5 before feed forward layers during training.

IV. EXPERIMENT RESULTS

A. EXPERIMENTAL DATASET AND EVALUATION METRICS

We conduct speaker-independent experiments using the IEMOCAP and MSP-IMPROV datasets. Both datasets are composed of multimodal interactions of dyadic sessions and labeled by three annotators for emotions such as happy, sad, angry, excited, and neutral, along with dimensional labels such as valence and arousal. In this study, we only use four emotional categories for both datasets: Happy, Sad, Angry, and Neutral.

The IEMOCAP dataset consists of five sessions, where each session contains scripted and improvised utterances from two speakers (one male and one female). For this study, we include excitement utterances with happiness ones. We take 5,531 utterances (1636 happy, 1084 sad, 1103 angry, 1708 neutral) for all sessions. The mean length of all the turns is 4.55 s (max.: 34.14 s, min.: 0.58 s).

The MSP-IMPROV dataset consists of six sessions in the same manner (12 unique speakers). Each session includes all the speaking turns of the improvisation and the natural interaction based on the 20 target sentences in the improvised scene. The final dataset contains a total of 7798 utterances (2644 happy, 885 sad, 792 angry, 3477 neutral). The mean length of all the turns is 4.09 s (max.: 31.09 s, min.: 0.41 s).

Since the input length for a CNN has to be equal for all samples, we set the maximal length to 7.5 s (mean duration plus standard deviation). Longer turns are cut at 7.5 s, and shorter ones are padded with zeros. The class distribution is unbalanced in both datasets, especially for MSP-IMPROV dataset, the number of utterances belonging to happy/neutral class more than three times that of angry/sad. Unweighted accuracy (UA) is the average classification accuracy for each emotion. It is a better measurement if the class distribution is not balanced. Hence, we use UA as the performance metric of the proposed framework to avoid being biased to the larger classes.

TABLE 2. Accuracy comparison of static features on IEMOCAP and MSP-IMPROV dataset (%).

Static features	UA	
	IEMOCAP	MSP-IMPROV
IS09	53.4	41.2
emobase2010	54.9	40.9
IS13 ComParE	54.5	40.6
MFCC	51.5	40.5
MSF	52.5	43.2

B. EMOTION RECOGNITION SYSTEM WITH STATIC FEATURES

Firstly, we investigate the conventional emotion recognition system with static features which are computed using fixed statistical functions to the hand-crafted LLDs. We extract MFCC, emobase2010, IS09 [36], and IS13 ComParE [37] features using the Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) toolkit [38]. All features are first normalized by specific z-normalization. Secondly, to investigate the effectiveness of static modulation features on emotion recognition, we also extract the MSFs by calculating the spectral centroid, spread, skewness, and kurtosis from the modulation spectral representation. For each feature set, we train a linear SVM model to recognize the speech emotion using LibSVM [39] and Weka toolkits [40]. All results are presented by leave-one-session-out cross-validation. Table 2 shows the accuracy comparison of static features on IEMOCAP and MSP-IMPROV datasets. The best result is 54.9 percent for IEMOCAP using the original static features with 1,582 dimensions whereas the best result is 43.2 percent for MSP-IMPROV using the static modulation features with 160 dimensions. The results also show that MFCC features achieve the worst results, which may be due to the minimum number of MFCC features (only 39 dimensions features). Similar to the results from [6], the MSFs perform better than MFCC for emotion recognition on both datasets. Emotion information from speech changes dynamically over time, but the static features do not contain temporal dynamics information which plays a key role in the emotion recognition process.

C. SETUP OF AUDITORY-BASED DEEP LEARNING MODELS

In the front-end signal processing, we first resample the speech signal with a sampling frequency of 16000 Hz and apply a pre-emphasis filter to compensate for the effect of sound source. We subsequently use normalization to remove the difference of the speakers by mapping the signal values to mean 0 and the standard derivation to 1 in each utterance. The sound-pressure level is set to 60 dB, which approximates to a normal voice. Furthermore, we introduce the compressive Gammachirp filterbank with 32 filters to provide the compressive characteristics. The frequency of Gammachirp filterbank distributed on the ERB_N scales is between 0.1 and 8 kHz. The modulation filterbank is also used to control the envelopes of octave bands from 2 to 512 Hz, consisting of

nine filters (one low-pass filter and eight band-pass filters). The low-pass filter is a 2nd order Butterworth infinite impulse response (IIR) filter with a cut-off frequency of 2 Hz. The cut-off frequencies of the band-pass filters are equally spaced on a logarithm scale from 2 to 512 Hz.

In the back-ends of the SER system, a joint deep learning model combined 3D convolution and ASRNN is used. To train the model with a speaker-independent property, we use leave-one-session-out cross-validation. In each experiment, four sessions are used for training the deep model and one session is divided into two sub-sessions depending on the gender in both datasets. For all random weight initializations, we choose L2 regularization. The parameters are learned in an end-to-end manner, meaning that all parameters of the model are optimized simultaneously using the Adam optimization method with a learning rate of 1e-4 to minimize cross-entropy loss. The batch size is 10, and maximum epoch is 30 with early-stopping. The process stops if the UA does not improve for 8 consecutive epochs.

D. IMPACTS OF SLIDING WINDOW AND SHIFT LENGTHS

SRNNs are used to obtain continuous internal representations while maintaining good computational efficiency. The continuous internal representations can be extracted using a sliding window. At the same time, computational efficiency can be improved by segmenting a feature sequence into multi sub-sequence. However, choosing different lengths of window and shift will affect the recognition accuracy and computational efficiency of emotional recognition system.

To reach higher recognition accuracy and computational efficiency, we investigate the effect of the sliding window and shift lengths using IEMOCAP dataset. First, the entire feature sequence is divided into multi-subsequences in a sliding manner. The length of each subsequence is much shorter than the original sequence, and the model can be trained rapidly using BPTT. Then, we run the proposed system five times and obtain the average accuracy in the case of different sliding window and shift lengths. We consider the different sliding window lengths of 10, 20, 30, 40, 50, and 100, which mean the duration of the sequence from 200 to 2000 ms. We also consider the shift lengths of 5, 10, and 20, which mean that it will produce 150, 75, and 38 sliding subsequences in the same padding manner for the duration of the convolutional sequence with 750×512. When the sliding window length is 100 with a shift length of 10, the training time of the ASRNN architecture is close to that of the entire sequence fed into the recurrent networks. Hence, we do not consider a longer sliding window that will take longer time in training the model. One session in the dataset is chosen for testing and others for training. We find that the computational efficiency will be improved with the shortening of window length and the lengthening of shift. But in this case, the recognition accuracy will decrease due to the inability to extract more emotional features. In addition, because only the feature of the last time frame in each sliding window is retained, when the window length is too long, not only the computational efficiency will be reduced, but also the recognition accuracy will be reduced. The results obtained for each method are shown in Fig.5. Recognition accuracy is closer when the shift length is 5 or 10, but it became worse when the shift length is 20. This figure also shows that the ASRNN architecture resulted in better accuracy when the sliding window length is 20 or 40. Therefore, we only consider sliding window lengths of 20 and 40 and shift lengths of 5 and 10.

TABLE 3. Accuracy comparison with different sliding-widows and shift lengths in ASRNN architecture on IEMOCAP and MSP-IMPROV dataset (%).

Sliding window length	Shift length	UA	
		IEMOCAP	MSP-IMPROV
20	5	62.3	54.9
20	10	62.6	55.7
40	5	61.0	54.2
40	10	62.1	55.3

TABLE 4. Confusion matrix (%) of ASRNN with an average accuracy of 62.6% on the IEMOCAP dataset.

		Output			
	Emotion	Neutral	Happiness	Anger	Sadness
Input	Neutral	58.5	17.0	8.0	16.5
	Happiness	20.6	55.6	12.7	11.1
	Anger	12.9	18.1	64.4	4.6
	Sadness	15.6	9.4	3.0	72.0

TABLE 5. Confusion matrix (%) of ASRNN with an average accuracy of 55.7% on the MSP-IMPROV dataset.

		Output			
	Emotion	Neutral	Happiness	Anger	Sadness
Input	Neutral	45.0	34.8	8.6	11.6
	Happiness	17.5	67.5	10.3	4.7
	Anger	13.2	25.1	59.7	2.0
	Sadness	22.5	23.3	3.7	50.5

E. RESULTS WITH ASRNNS ARCHITECTURE

Table 3 shows the recognition results using different lengths and shift of the sliding window with the ASRNNs architecture for both datasets. One can see that the ASRNNs architecture with the sliding window length of 20 and shift length of 10 performed better than the others, whose recognition accuracy is 62.6% for IEMOCAP and 55.7% for MSP-IMPROV. The results are much better than those obtained using the traditional parameters shown in Table 2. According to the results, the window length of 20 frames (about 400ms) is suitable for expressing segment-level emotions, while the shift length of 10 is better for classification than that with the shift length of 5. Comparing with the best results of traditional recognition system in Table 2, the proposed system achieved $+7.7$ and $+12.5\%$ absolute accuracy improvements on IEMOCAP and MSP-IMPROV, respectively. These results indicate that the proposed system with temporal dynamics information is better to recognize emotional states than the conventional system with static features.

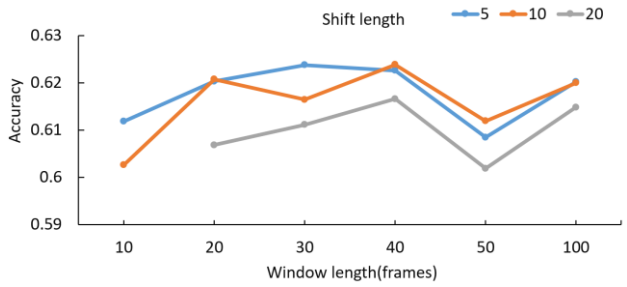

FIGURE 5. The impact of sliding window and shift length on recognition accuracy.

Table 4 and 5 show the confusion matrix of the best results for the IEMOCAP and MSP-IMPROV datasets, respectively. In general, the class distributions of the confusion matrix for different session are basically similar. One can see that happiness is easily confused with neutral emotion and vice versa. Anger is more easily misclassified as happiness than happiness being misclassified as anger. Unlike the study [41], the proposed system reduces the confusion between anger and happiness categories to a major extent, especially in MSP-IMPROV. Sadness is easily confused with neutral emotion in IEMOCAP, while it is easily confused with happiness in MSP-IMPROV. The confusion in the proposed method mainly happen between the neutral one and the others. This implies that emotion recognition based on auditory front-ends is basically consistent with people's recognition of emotion. In terms of the databases, the overall performance on IEMOCAP is better than MSP-IMPROV. The reason for this seems to be that the MSP-IMPROV dataset is highly imbalanced.

F. IMPACTS OF MODULATION CHANNEL, SLIDING WINDOW AND ATTENTION MODEL

In order to evaluate the effects of modulation channel number, sliding window and attention model on the SER system, we design a number of comparative experiments in different situations.

First, we evaluate the effects of the nine modulation filterbank in obtaining local features and periodicity information by comparing it to the one with six modulation filters (ASRNN-6MFB). ASRNN-6MFB is set as the same layers as the ASRNN, but different inputs shape of $32 \times 6 \times 6000$ result in different kernel and stride. Compare to ASRNN, the difference is that the kernel and stride are $2 \times 1 \times 2$ instead of $2 \times 2 \times 2$ in Pool2. In addition, the convolutional maps are 40 instead of 64 to keep similar features in each frame. Finally, the output shape is $4 \times 3 \times 750$ in pool3. Then this layer is reshaped to 2D shapes of 750×480.

Second, an attention-based recurrent neural network (ARNN) is designed to evaluate whether the sliding window can obtain more temporal dynamics information or not. ARNN is a special case of an ASRNN. That is, when the sliding window length of an ASRNN is equal to the length of the entire convolution sequence and the shift length is equal to 0, it becomes an ARNN. Hence, the attention model is used on the entire time sequence.

TABLE 6. Accuracy comparison (%) between RNN architectures on the IEMOCAP and MSP-IMPROV dataset.

RNN architecture	UA	
	IEMOCAP	MSP-IMPROV
SRNN-Max-pooling	61.5	54.2
SRNN-Mean-pooling	61.7	53.9
ARNN	61.3	55.2
ASRNN-6MFB	61.7	54.8
ASRNN	62.6	55.7

Third, SRNNs with max and mean pooling are designed to evaluate whether the attention model can seize the emotional regions. A SRNN has the same sliding window and shift lengths as the ASRNN. There are two types of pooling used in a SRNN: maximum and average, denoted as SRNN-Max-pooling and SRNN-Mean-pooling, respectively. These models mentioned above use the same convolutional networks with the input shape of $32 \times 9 \times 6000$.

Table 6 shows the comparison of results on different types of SRNNs with attention and non-attention models and one ARNN. Compared with ASRNN-6MFB, the ASRNN achieves the same improvements of $+0.9\%$ on both datasets. This means that the proposed system with nine channels may extract more information from speech than ASRNN-6MFB. Compared with the ARNN, the ASRNN achieves $+1.3\%$ and $+0.5\%$ absolute improvements on the IEMOCAP and MSP-IMPROV datasets, respectively. This means that the segment-based attention model is better than frame-based attention model. Compared with SRNN-Max-pooling and SRNN-Mean-pooling, the ASRNN achieves $+0.9\%$ and $+1.5\%$ absolute improvements on the IEMOCAP and MSP-IMPROV datasets, respectively. This means that the attention model is better than max- and mean-pooling.

G. LISTENING TEST FOR TEMPORAL ATTENTION

Recently, Kell *et al.* [42] demonstrated that a deep neural network made human-like error patterns. If our attention model reflects human mechanism, its result should be similar to human behaviors when they recognize speech emotion. For this reason, a listening testing is designed to evaluate the similarity of the behaviors between the proposed attention model and human. Thirty sentences from IEMOCAP dataset are used for the listening tests. Each sentence with a duration between 4.5 to 7.5 s is presented to at least 25 listeners (14 female and 11 male with ages ranging from 20 to 28) in random orders. The listeners are asked to concentrate on listening to each utterance and choose the two locations that best show the emotions of the utterance.

Figure 6 illustrates an example of comparisons between the attention model and human temporal attention. The top panel shows the waveform of an emotional sentence, and the upper middle panel shows the spectrogram of the sentence. The lower middle panel shows the attention weights (α_i) that are calculated based on auditory front-ends and deep

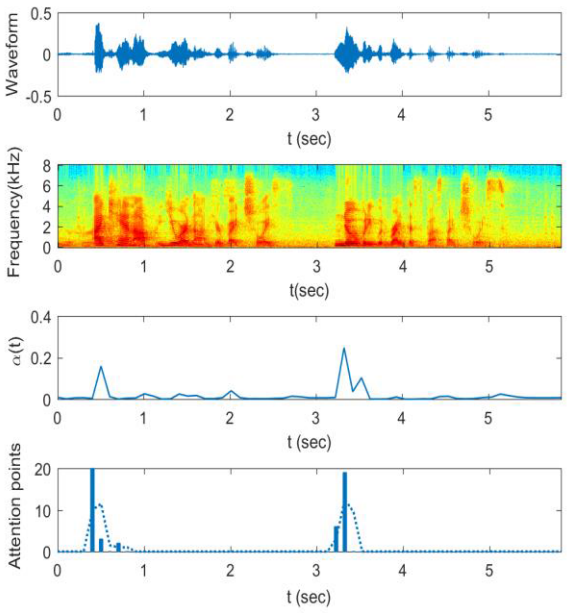

FIGURE 6. Analysis and comparison of attention model and human selective attention for test example. Top panel: raw waveform (Ses01F_impro04_F033.wav from IEMOCAP dataset); upper middle panel: spectrogram; lower middle panel: attention weight (α_i) over sliding window time sequence; bottom panel: histogram shows attention numbers for subjective judgments, and dashed line shows moving-average with 2 data points.

frameworks. The bottom panel shows a histogram that is the point numbers of attention position given by subjective judgements, and a dashed line that is the moving-average on two neighbor data points. One can see that the curve of the attention weights is similar to that of the subjective judgment. Pearson's correlation coefficient is used to quantitatively measure the similarity between the attention model and human temporal attention. The correlation coefficient is $P = 0.552$ ($\rho < 0.001$) between the attention weights and histogram in this particular utterance. If we calculate the correlation between the moving average values and the attention weights, the correlation coefficient becomes $P = 0.715$ ($\rho < 0.001$). This indicates that there is a strong correlation between human temporal attention and the attention model. This implies that the proposed attention model can reflect human selective attention to a large extent.

V. DISCUSSION

Taking into account, that the human auditory system has a very strong ability to perceive the intensity and fundamental frequency of speech, furthermore, it can track temporal dynamics of emotion from the perceived information and focus on the salient emotion regions, therefore, we propose a SER system by combining auditory mechanism and attention mechanism of human auditory system.

The auditory front-ends of the SER system are used to produce temporal modulation cues, which contain local features and periodicity information of emotional speech. During the process of temporal modulation cues extraction, an additional correlation in neighboring channels will be introduced because of the partially overlapped frequency. Traditional methods use discrete cosine transform to de-correlate the temporal modulation features in the acoustic and modulation frequency domains. Since CNNs can successfully de-correlate the features in neighboring channels, we directly use 3D CNNs to learn a joint spectral-temporal feature from temporal modulation cues. Furthermore, temporal dynamic information is obtained by continuously scanning the temporal sequence and then is transmitted to higher-level processing center. To focus on the emotional regions while ignore the emotionless regions, an attention model is used to extract utterance-level features.

TABLE 7. Accuracy comparison of proposed system and other systems on IEMOCAP and MSP-IMPROV dataset (%).

Literature	Features	Backend	UA	
			IEMOCAP	MSP-IMPROV
Ref [43]	Raw speech	CRNN	60.23	52.43
Ref [41]	Log Mel-filterbank	Attentive CNN	59.54	45.76
Ref [44]	Mel-filterbank	CNN	61.8	53.8
Ref [45]	LLDs	Deep belief network	62.4	-
Ref [46]	FFT bins	BLSTM	52.8	-
Ref [47]	LLDs	Attention-based BLSTM	60.1	-
Ref [10]	Temporal modulation	3D-CRNN	60.93	-
Proposed	**Temporal modulation**	**ASRNN**	**62.6**	**55.7**

To show the benefit of the proposed model, we compare our results with the studies [41], [43], [44] on both datasets as presented in Table 7. In [43], the authors used Mel filterbank features as the input to CNNs and showed that CNNs with these features can produce competitive results to the popular feature sets. In [41], the authors used Log-Mel filterbank features as the input to autoencoder and used attentive CNN for representation learning. In [44], the authors used raw speech as input to parallel convolution layer and showed that CNN-LSTM can capture multi-temporal dependencies. Compared to these studies, we are achieving the better result of 62.6% and 55.7% respectively on both datasets using 3D convolutions and ASRNNs from temporal modulation cues. This indicates that the auditory front-ends can provide spectral-temporal representations, and deep frameworks can effectively extract emotional information from such representation for emotion recognition.

In addition, four representative studies with reported results on IEMOCAP are selected as comparisons. In [45], the authors used static features of LLDs for representation learning, and deep belief network for emotion recognition. In [46], the authors used FFT bins with autoencoder for representation learning, and used RNN to identify the emotion states. In [47], the authors used attention-based BLSTM models on LLDs for emotion recognition. Additionally, compared with our previous study [10], we are able to obtain faster

training speed with SRNNs, and this system can better identify happiness and anger. This may be benefited by the 9-channel modulation filterbanks that contain fundamental frequency information, which is important for emotions. In contrast, our study exceeded the accuracy compared to the leading studies.

Other studies used attention models to identify emotions on IEMOCAP databases, but the experimental conditions are different. For example, [25], [29], [30] did not merge happy and excited into one class, while [33] just reported weighted accuracy. Unlike these frame-based attention models, we use a sliding window based attention model to focus on the salient emotion regions. The results of experiments showed that this model can effectively obtain the emotional information. The subjective evaluation shows that the attention patterns of the attention model are basically consistent with human behaviors in recognizing emotions.

VI. CONCLUSION

We proposed a SER system using 3D convolutions and attention-based sliding recurrent neural network based on auditory front-ends. As the human auditory system is powerful in spectral-temporal signal analysis and processing, an auditory model, which mimics the function of the human auditory system, is used as a front-end to extract spectral-temporal features in the SER system. Additionally, compared with modulation spectral features, these 3D features contain temporal dynamics characteristics and can avoid the modulation correlation problem.

Considering that local features and periodicity information can better express emotions, we used 3D convolutions to extract frame-level features from nine modulation filters. We then used recurrent networks to obtain temporal dynamics information in each utterance. We also used an attention model to focus on the emotionally salient parts of a speech signal. Therefore, we propose a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks. To the best of our knowledge, this is the first study on speech emotion recognition combining auditory and cognitive mechanisms. Our experiments demonstrated that the proposed system can obtain spectral-temporal representations and exhibit better recognition accuracy compared to that of state-of-the-art SER systems on both datasets.

In summary, an auditory model as a front-end can extract rich spectral-temporal information, and the proposed system can effectively extract high-level features for emotion recognition. This system is possibly applied to other audio-event perception and recognition. For future work, we plan to conduct an experiment using categorical and dimensional speech emotional datasets to analyze noise-robust emotion recognition. In addition, inspired from the study [48] using a filterbank layer in DNN to learning the filterbank features, we plan to design the auditory and modulation filterbank layers to produce 3D spectral-temporal representations for emotion recognition.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.
[2] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 7, pp. 668–675, 2003.
[3] J. Mcdermott and E. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, Sep. 2011.
[4] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 39, no. 6, pp. 379–386, Nov. 2018.
[5] N. Moritz, J. Anemueller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1926–1937, Nov. 2015.
[6] 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, May 2011.
[7] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," in *Proc. Interspeech*, 2016, pp. 262–266.
[8] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 39, no. 3, pp. 234–242, May 2018.
[9] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2719–2732, Nov. 1999.
[10] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on spectral-temporal representation," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2018, pp. 1–6.
[11] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 336, no. 1278, pp. 367–373, 1992.
[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
[13] C. Busso, S. Parthasarathy, A. Burmania, M. Abdelwahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.
[14] T. Irino and R. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 2222–2232, Nov. 2006.
[15] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *J. Acoust. Soc. Amer.*, vol. 114, no. 3, pp. 1529–1542, Sep. 2003.
[16] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore, and R. D. Patterson, "Comparison of the roex and gammachirp filters as representations of the auditory filter," *J. Acoust. Soc. Amer.*, vol. 120, no. 3, pp. 1474–1492, Sep. 2006.
[17] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Amer.*, vol. 102, no. 5, pp. 2906–2919, Nov. 1997.
[18] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, Sep. 1983.
[19] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, nos. 1–2, pp. 103–138, Aug. 1990.
[20] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, p. 3009, May 1994.
[21] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[22] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

[23] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.
[24] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3412–3419.
[25] M. 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.
[26] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 383–388.
[27] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.
[28] C. Stevens and D. Bavelier, "The role of selective attention on academic foundations: A cognitive neuroscience perspective," *Develop. Cognit. Neurosci.*, vol. 2, pp. S30–S48, Feb. 2012.
[29] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.
[30] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. Interspeech*, Aug. 2016, pp. 1387–1391.
[31] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5150–5154.
[32] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 1537–1540.
[33] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. Interspeech*, no. 3, Aug. 2017, pp. 1263–1267.
[34] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, Sep. 2019, pp. 2803–2807.
[35] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2578–2582.
[36] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2009.
[37] B. Schuller, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Aug. 2013, pp. 1–5.
[38] 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*, 2010, pp. 1459–1462.
[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[41] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7390–7394.
[42] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. Mcdermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630.e16–644.e16, May 2018.
[43] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," 2019, *arXiv:1904.03833*. [Online]. Available: https://arxiv.org/abs/1904.03833
[44] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, no. 3, Mar. 2017, pp. 2741–2745.
[45] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, Jan. 2017.
[46] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. Interspeech*, 2016, pp. 3603–3607.
[47] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. Interspeech*, Aug. 2018, pp. 272–276.
[48] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 297–302.

ZHICHAO PENG received the B.S. degree in computer science from Hunan Normal University, China, in 2000, and the M.S. degree in signal and information processing from Central South University, China, in 2007. He is currently pursuing the joint Ph.D. degree with the Japan Advanced Institute of Science and Technology, Japan, and Tianjin University, China. His research interests include emotion recognition, deep learning, auditory signal processing, and speech and language processing.

XINGFENG LI received the B.E. degree in software engineering from the Changchun University of Science and Technology, China, in 2013, and the M.S. degrees in software engineering and information science from Tianjin University, China, and Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2016. He started his research as a member at the Acoustic Information Science (AIS) Laboratory, JAIST, in 2014. His research interests are in affective computing, speech processing, and speech perception with an emphasis on how para/non-linguistic information (speech emotion) impacts spoken communication.

ZHI ZHU received the B.E. degree in communication engineering from the Nanjing University of Posts and Telecommunications, in 2012, and the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology, in 2015 and 2018, respectively. He is currently a Scientist with Fairy Devices Inc. He is interested in hearing and speech science.

MASASHI UNOKI received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. His main research interests are in auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He was a Visiting Researcher with the ATR Human Information Processing Laboratories, from 1999 to 2000 and the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been on the faculty of the School of Information Science, JAIST, since 2001, where he is currently a Full Professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). He received the Sato Prize from ASJ, in 1999, 2010, and 2013 for an Outstanding Paper and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.

JIANWU DANG received the B.E. and M.E. degrees from Tsinghua University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Japan, in 1992. He was a Lecturer with Tianjin University, Tianjin, China, from 1984 to 1988. From 1992 to 2001, he worked at ATR Human Information Processing Laboratories, Japan. Since 2001, he has been on the faculty of the School of Information Science of JAIST as a Professor. He joined the Center of National Research Scientific, Institute of Communication Parlee (ICP), France, as a Research Scientist the first class, from 2002 to 2003. Since 2009, he has been with Tianjin University. He built a 3D physiological model for speech and swallowing, and endeavors to apply the model on clinics. His research interests include speech production, speech synthesis, and speech cognition.

MASATO AKAGI received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been on the faculty of the School of Information Science, JAIST, where he is currently a Full Professor. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech.