Spectro-Temporal Modulations Incorporated Two-Stream Robust Speech Emotion Recognition

Yih-Liang Shen, Pei-Chin Hsieh, and Tai-Shih Chi

***Abstract*—Deep learning based speech emotion recognition (SER) models have shown impressive results in controlled environments, but their performance significantly degrades in noisy conditions. This paper proposes a robust two-stream SER model by combining spectro-temporal modulation features with conventional acoustic features. Experiments were conducted on German (EMODB) and English (RAVDESS) datasets using the clean-train-noisy-test paradigm. The results demonstrate that spectro-temporal modulation features offer superior robustness in noisy conditions compared with conventional acoustic features such as MFCCs and time-frequency features from Mel-spectrograms. Additionally, we analyze weights of modulation features and demonstrate the model emphasizes contours of formants and harmonics, which are crucial features for speech perception in noise, for robust SER. Incorporating the stream of spectro-temporal modulations not only enhances the robustness of the model but also provides deeper insights into the task of SER in noise.**

***Index Terms*—Speech emotion recognition, auditory model, spectral-temporal modulation.**

I. Introduction

SPEECH emotion recognition (SER) systems are designed to identify emotion by analyzing the emotional content in human speech. SER can enhance the user experience of human-computer interaction by allowing machines to respond appropriately to the user's emotional state. For instance, SER can be used in customer service management systems to analyze the emotional tone in customer calls. In summary, SER has diverse applications across various industries, contributing to improved human-computer interaction, customer service, and healthcare.

SER technology has been developed for a long time. Before the era of deep learning, researchers proposed various systems using different features and recognizers. The low-level features such as pitch and energy contour with Gaussian mixture models (GMMs) were proposed for SER in [1]. In [2], researchers used the fundamental frequency (F0), energy, zeros crossing rate (ZCR), linear predictive coding (LPC) features, and Mel-frequency cepstral coefficient (MFCCs) as a feature set with the support vector machine (SVM) recognizer. During the past decade, researchers have also proposed many SER methods based on neural networks (NNs). Some of them use NNs, such as multilayer perceptron (MLP) [3], convolution neural network (CNN) [4], and recurrent neural network (RNN) [5], as feature extractors. Some even integrate several types of neural networks into a system to exploit their architecture's characteristics [6].

These deep learning SER methods have achieved good performance in a laboratory environment, which is noise-free and has no reverberations. However, the real-world environment where the SER system is used contains background noise, such as outdoor traffic noise for mobile users, and indoor fan or TV noise for at-home users. Experiment results show that the existing SER systems are sensitive to background noise and degrade severely while facing noise [7], [8], [9]. Hence, developing a noise-robust SER system for realistic environments is important and has drawn researchers' attention. To enhance the robustness of the SER model, one can adopt data augmentations by including several types of noise at various signal-noise-ratio (SNR) conditions when training the model as in [10]. The other type of approach would be using less noise-degraded features as in [11], [12].

Falling into the approach of using less noise-degraded features, our previous study shows that the spectro-temporal features extracted from a multi-resolution auditory perceptual model can be adopted for a robust SER system with the SVM classifier [13]. This two-stage auditory perceptual model comprises a stage of simulating cochlear frequency analysis, and a following stage of simulating the spectro-temporal analysis of the auditory cortex (A1) in the brain. The output spectro-temporal features of the auditory model encode the spectral and temporal structures of speech jointly. In this paper, we examine our idea in the NN era by proposing using spectro-temporal auditory features with an attention-based convolutional RNN (ACRNN) model [14] for robust SER. We evaluate the model using two public SER datasets, EMODB and RAVDESS, under various noise conditions. The results demonstrate that spectro-temporal auditory features provide noise-robust information for speech emotion recognition in the clean-train-noisy-test scenario. Additionally, we propose a two-stream ACRNN (tACRNN) that simultaneously takes the Mel spectrogram and the spectro-temporal auditory features as input. The proposed tACRNN model achieves moderate performance under both clean and noisy conditions. The main contributions of this paper are summarized below:

1) We incorporate the spectro-temporal features extracted from an auditory model into NN-based robust SER models.

Received 12 July 2024; revised 27 November 2024; accepted 15 January 2025. Date of publication 20 January 2025; date of current version 15 September 2025. This work was supported by the Ministry of Science and Technology, Taiwan under Grant MOST 110-2221-E-A49-115-MY3. Recommended for acceptance by J. Shukla. *(Corresponding author: Yih-Liang Shen.)*

The authors are with the Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: yihliang.ee06@nycu.edu.tw; wren.ee12@nycu.edu.tw; tschi@nycu.edu.tw).

Digital Object Identifier 10.1109/TAFFC.2025.3531638

2) We demonstrate that spectro-temporal features are more robust to additive noise for SER than traditional acoustic features such as MFCC and Mel spectrogram in NN-based models.
3) We demonstrate that a well-trained SER model on spectro-temporal features focuses on the harmonic structure and harmonic contours, which have been shown strongly helpful to speech perception in noise [15].

The rest of this paper is organized as follows. Section II briefly reviews related works, including works in SER, works in robust SER, and the two-stage auditory model with spectro-temporal features. Section III introduces our proposed tACRNN model for SER. The experiment setups and datasets are described in Section IV. Section V discusses the experimental results, while Section VI addresses the implications and limitations of this work. Finally, Section VII concludes the paper.

II. Background and Related Works

A. Speech Emotion Recognition

The purpose of SER is to recognize the user's emotional state using speech signals. An automatic SER system can mainly be divided into a feature extraction module and a classifier. Understanding which speech features carry significant emotion information is crucial for feature extraction. Some researchers have focused on prosody features such as energy, duration, and pitch tracks [16]. Others have utilized acoustic features like audio spectrograms [17] and Mel frequency cepstral coefficients (MFCCs) [18]. More recently, self-supervised learning (SSL) models have garnered attention and shown very promising results across various downstream tasks. These SSL models have been leveraged for SER, and with features extracted by them demonstrating effectiveness in the emotion recognition task [19], [20].

Classifiers are broadly categorized into traditional and deep learning-based models. Traditional classifiers encompass techniques such as Gaussian Mixture Model (GMM) [21], [22], k-means clustering [23], Support Vector Machine (SVM) [2], and decision tree [24]. These techniques have been explored and documented in many studies, highlighting their applicability and effectiveness in various fields. Over the past decade, deep learning-based methods have demonstrated remarkable success across a wide range of tasks, including SER. Among these, Yuan et al. applied MFCCs as the input feature along with the multilayer perceptron (MLP) classifier [25]. Peng et al. introduced a multi-scale convolutional neural network (MSCNN) aiming to leverage both audio and textual data to develop a multimodal SER system [26]. Considering the sequential nature of speech, which suggests that only certain segments may carry significant emotion information, Atmaja et al. employed a combination of long short-term memory (LSTM) networks and attention mechanisms [27]. This approach aims to automatically identify and emphasize the segments which are most informative for emotion recognition. More recently, transformer [28] has drawn wide attention in SER. Lu et al. use frame-based and segment-based transformers to aggregate the local and global emotion information [29].

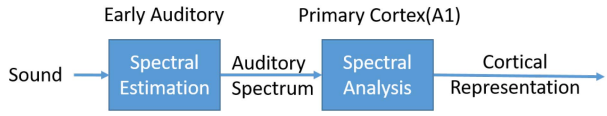

Fig. 1. Auditory model.

B. Robust Speech Emotion Recognition

While deep learning methods have demonstrated impressive performance on SER, the performance would be greatly deteriorated by ambient sounds, including background noise and reverberation. Several studies have explored strategies to mitigate the impact of noise on SER systems by emphasizing feature estimation on cleaned speech for enhanced performance. [30], [31], [32] have contributed to this area by employing denoising techniques aimed at predicting clean speech features, which are crucial for accurate SER. Another type of studies enhanced the noise resilience of SER systems through data augmentation and demonstrated the effectiveness in improving system robustness against various acoustic disturbances [33], [34]. However, these systems often experience significant degradation when encountering unseen disturbances.

Alternatively, some research endeavors have focused on identifying features that encode emotion information and exhibit robustness to noise. Leem et al. investigated the noise robustness of low-level descriptors (LLDs) features and showed that using noise-robust LLDs features for SER under noisy conditions is better than using all the LLDs features [11]. Another study showed that the temporal modulation spectrum contains valuable emotion state information [35] and the subsequent study highlighted the noise robustness of the temporal modulation spectrum features [36]. Moreover, incorporating a bag-of-word strategy on temporal modulation spectrum features was proposed to further improve SER performance in challenging 'in-the-wild' conditions [12].

C. Auditory Model

The auditory model, illustrated in Fig. 1 and established from physiological studies [37], [38], can be divided into two main parts: the early auditory (ear) module and the primary cortex (A1) module.

1) Early Auditory Module: The module mainly imitates the function of the cochlea, transforming sound waves into electrical signals to the auditory nerves. Because of the non-uniform texture of the basilar membrane, sound waves with different frequencies will cause the maximum vibration at different locations along the basilar membrane. Functionally speaking, we can view the cochlea as a frequency analyzer on the sound wave. In the original auditory model [39], this module is implemented by a band of 128 constant-Q bandpass filters. In this paper, rather than using the original auditory model, we use the logarithm power of the Mel-spectrogram to mimic similar frequency analysis on sound waves.

2) Primary Cortex Module and Spectro-Temporal Modulation Features: Neurons of the auditory cortex (A1) analyze the

auditory spectrogram generated from the early auditory module in the joint time-frequency domain. Animal experiments have found that A1 neurons respond strongly to sounds with specific time-frequency structures [38]. Therefore, A1 neurons can be modelled as 2D spectro-temporal modulation filters capturing different time-frequency structures of the sound. In [39], we implemented 2D filters, which tune to various spectro-temporal modulation parameters, to simulate the functions of the A1 neurons. The rate parameter ω (in Hz) describes how fast the instantaneous power of the auditory spectrogram varies over time. The scale parameter Ω (in cycle/octave) defines the width of the repetition span of the auditory spectrogram along the frequency axis. Positive and negative signs of the rate parameter represent the time-frequency structure with the downward and upward sweeping directions, respectively. Fig. 2(c) shows the spectro-temporal impulse response of a downward-sweeping modulation filter. These modulation filters are applied to the output of the early auditory module as follows:

$$r(t, f, \omega, \Omega) = x(t, f) *_{tf} \mathrm{M}(t, f; \omega, \Omega) \tag{1}$$

where $x(t, f)$ is the output of the early auditory module, $\mathrm{M}(t, f; \omega, \Omega)$ is the impulse response of the modulation filter tuned by parameter ω and Ω. $*_{tf}$ is the 2D convolution operating in the time-frequency domain. More detailed descriptions can be accessed in [40].

In this study, we apply these spectro-temporal modulation filters on the Mel-spectorgram, which imitates the auditory spectrogram generated by the early auditory module, to produce the spectro-temporal modulation features. After applying modulation filters, the output can be written as

$$C(n, k; \pm\omega_0, \Omega_0) = \mathcal{F}_{2D}^{-1}\{\mathcal{F}_{2D}\{X(n, k)\} \cdot STMF(\pm\omega_0, \Omega_0)\} \tag{2}$$

where $X(n, k)$ is logarithm of Mel-spectrogram, n and k are time and frequency indexes, respectively. $\mathcal{F}_{2D}$ and $\mathcal{F}_{2D}^{-1}$ are the 2D Fourier transform and the inverse 2D Fourier transform. $STMF(\pm\omega_0, \Omega_0)$ is the frequency response of a spectro-temporal modulation filter tuned to rate $= \omega_0$ and scale $= \Omega_0$. $STMFs$ of the positive and negative rates are formulated as (3) and (4) shown at the bottom of this page, where $\mathcal{F}$ is the 1D Fourier transform; $\otimes$ is the outer product; π is half of the sampling frequency along the time and the frequency axes; $h_{scale}(k)$ and $h_{rate}(n)$ are 1D spectral and temporal impulse responses of the modulation filter. These 1D impulse responses are derived as the sinusoidal modulated gamma distribution

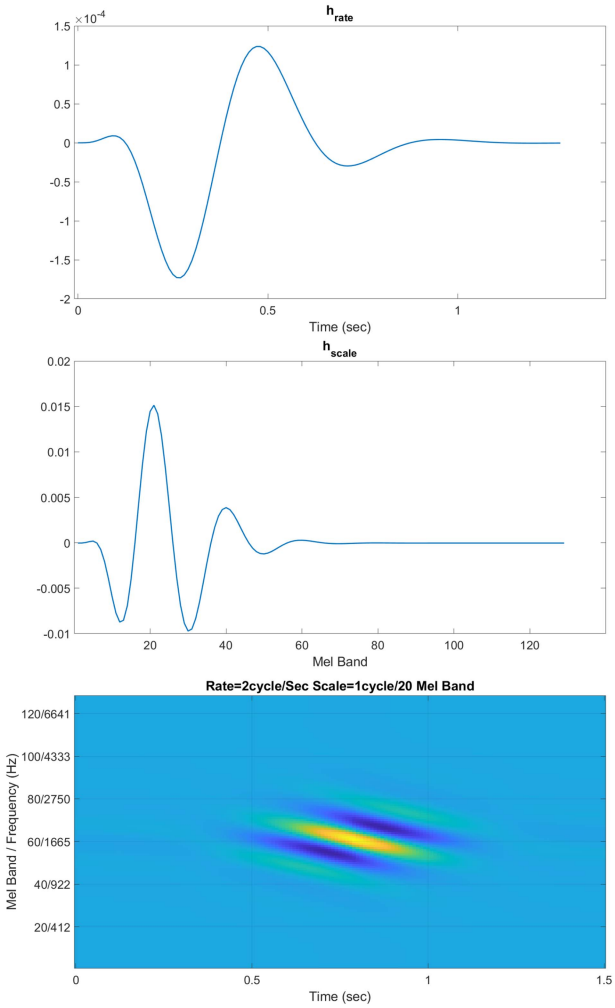

Fig. 2. Impulse response of a spectro-temporal modulation filter. (a) Top: h_{rate} tuned to $\omega = 2$ Hz. (b) Middle: h_{scale} tuned to $\Omega = 1$ cycle/20 Mel-bands. (c) Bottom: Impulse response of the downward spectro-temporal modulation filter formed by h_{rate} and h_{scale}.

functions as follows:

$$\begin{cases} h_{rate}(n; \omega_0) = \hat{n}^4 e^{-2\pi B_{rate}\hat{n}} \cos(2\pi\omega_0\hat{n}), \\ h_{scale}(k; \Omega_0) = \hat{k}^4 e^{-2\pi B_{scale}\hat{k}} \cos\left(2\pi\Omega_0\hat{k}\right), \end{cases} \tag{5}$$

where $\hat{n} = n/\#$ of frames per second; $\hat{k} = k/20$. Consequently, the units of rate (ω) and scale (Ω) are cycle/second (Hz) and cycle/20 Mel-bands. Bandwidths B_{rate} and B_{scale} are increased with the center frequencies ω_0 and Ω_0, respectively. That is, h_{rate} and h_{scale} are constant-Q bandpass filters. In this study, we

$$STMF(+\omega_0, \Omega_0) = \begin{cases} |\mathcal{F}\{h_{rate}(n; \omega_0)\} \otimes \mathcal{F}\{h_{scale}(k; \Omega_0)\}|, & 0 \le \omega < \pi; 0 \le \Omega < \pi \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

$$STMF(-\omega_0, \Omega_0) = \begin{cases} |\mathcal{F}\{h_{rate}(n; \omega_0)\} \otimes \mathcal{F}\{h_{scale}(k; \Omega_0)\}|, & -\pi < \omega \le 0; 0 \le \Omega < \pi \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

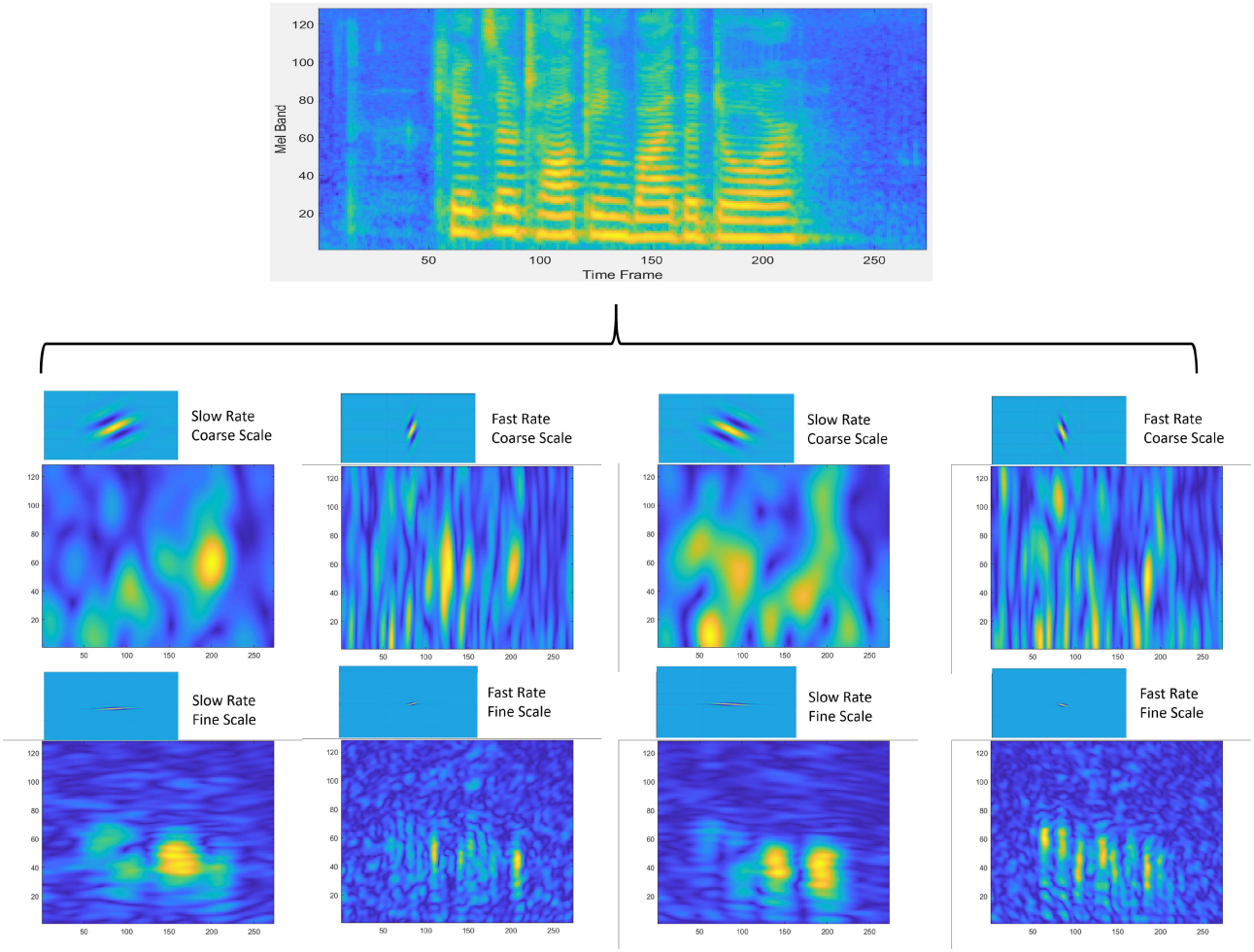

Fig. 3. Spectro-temporal modulation features: instantaneous power of the outputs of the spectro-temporal modulation filters.

used $Q_{3\,\mathrm{dB}} \approx 1.3$ and 0.4 in ω and Ω dimensions. We set $Q^{\omega}_{3\,\mathrm{dB}} \approx 1.3$ based on the parameter setting in [36], and $Q^{\Omega}_{3\,\mathrm{dB}} \approx 0.4$ based on the parameter setting in our previous work [41]. We tried several $Q_{3\,\mathrm{dB}}$ settings and found out these settings produced the optimal results. Hence, we used these settings in all experiments.

Fig. 2(a)(b) show h_{rate} and h_{scale} tuned to $\omega = 2$ Hz and $\Omega = 1$ cycle/20 Mel-bands. Fig. 3 shows some samples of spectro-temporal modulation features. A Mel-spectrogram of a sample utterance is shown in the top panel. There are 8 sets of subfigures, corresponding to 8 spectro-temporal modulation filters, shown at the bottom. The impulse response of each filter is plotted in the smaller subfigure while the instantaneous power of the output of the filter is plotted in the larger subfigure. Subfigures in the left/right two columns correspond to modulation filters with the upwards/downward sweeping direction. In this paper, we use the instantaneous power of the output of the filters as the spectro-temporal modulation features.

III. Proposed SER Model

We incorporate the spectro-temporal modulation features with the commonly-used time-frequency representation, log Mel-spectrogram, to provide complementary information for robust SER. Fig. 4 shows the proposed two-stream model. The overall model is extended from the attention-based convolutional RNN (ACRNN) [14] by adding a modulation branch. The modulation branch extracts features from spectro-temporal modulations, and extracted modulation features are concatenated with the output of the Mel-spectrogram branch for robust SER. Finally, the recognition branch predicts the emotional state based on the information from the previous two branches.

A. Spectrogram Branch

This branch takes log Mel-spectrogram as the input and uses a series of convolutional layers (Conv1, Conv2,..., ConvN) with varying kernel sizes, strides, and numbers of kernels. The convolutional layers are followed by max-pooling operations to reduce the dimensionality and fully connected layers to consolidate features. We integrate batch normalization and LeakyReLU activation functions to enhance training efficiency and non-linear representation. We have tried using five convolutional layers as in the original ACRNN paper [14]. However, the performance difference between using five convolutional layers and two convolutional layers is only less than 1% under the clean condition. Thus, we only use two convolutional layers in

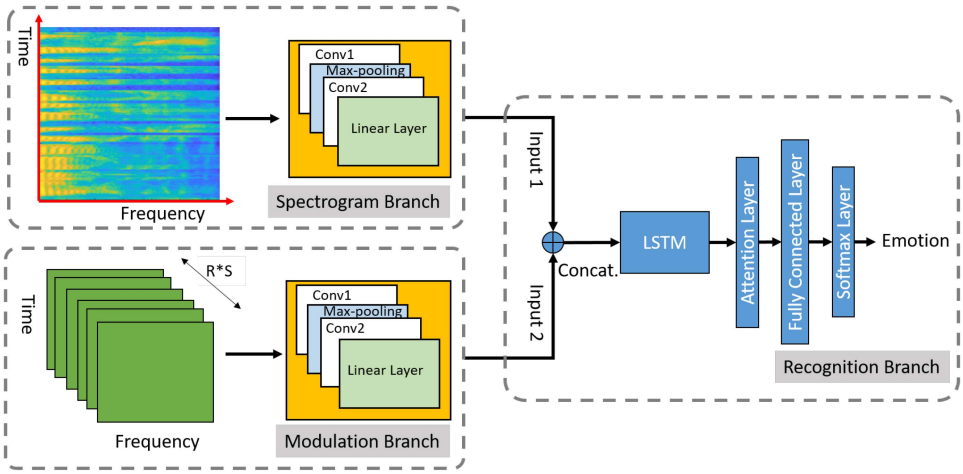

Fig. 4. Proposed two-stream robust SER model.

TABLE I
ARCHITECTURE OF THE SPECTROGRAM BRANCH IN FIG. 4

Layer Type	Kernel Size	Stride	# of Kernels	Output Size
Input	-	-	-	1*T*128
CONV + LRLU	5X3	1X1	128	128*T*128
MaxPool	2X4	2X4	-	128*(T/2)*32
Dropout	-	-	-	128*(T/2)*32
CONV + LRLU	5X3	1X1	256	256*(T/2)*32
Dropout	-	-	-	256*(T/2)*32
FC	8192X768	-	-	(T/2)*768
BN + LRLU	-	-	-	(T/2)*768

"CONV" represents a convolutional layer, "LRLU" is short for LeakyReLU, "FC" represents a fully connected layer, and "BN" is short for batch normalization.

TABLE II
ARCHITECTURE OF THE MODULATION BRANCH IN FIG. 4

Layer Type	Kernel Size	Stride	# of Kernels	Output Size
Input	-	-	-	40*T*128
CONV + LRLU	5X3	1X1	128	128*T*128
MaxPool	2X4	2X4	-	128*(T/2)*32
Dropout	-	-	-	128*(T/2)*32
CONV + LRLU	5X3	1X1	256	256*(T/2)*32
Dropout	-	-	-	256*(T/2)*32
FC	8192X768	-	-	(T/2)*768
BN + LRLU	-	-	-	(T/2)*768

See Table I for abbreviations.

ACRNN architectures in this paper to save training time. Parameters of the two convolutional layers are specified in Table I.

B. Modulation Branch

Detailed in Table II, this branch operates on spectro-temporal modulation features. The model architecture in this branch is similar to the architecture in the spectrogram branch. The ACRNN architecture has been demonstrated to effectively extract information from log Mel-spectrograms [14], while spectro-temporal modulation features can be considered as filtered spectrograms. Therefore, we adopt a similar architecture for the modulation branch in this work. The input size is (R*S)*T*128, where R is the number of chosen rate parameters, S is the number of chosen scale parameters, T is the number of time frames, and 128 is the number of Mel-frequency bins.

TABLE III
ARCHITECTURE OF THE RECOGNITION BRANCH IN FIG. 4

Layer Type	Output Size
Input 1	(T/2)*768
Input 2	(T/2)*768
Concatenate	(T/2)*1536
BLSTM (128 hidden units)	(T/2)*256
Attention on Time	256
FC	64
LRLU + Dropout	64
FC	# of class

See Table I for abbreviations.

C. Recognition Branch

As shown in Table III, the extracted features from the spectrogram and modulation branches are concatenated first, and sent to a bi-directional long short-term memory (BLSTM) layer with 128 hidden units, which processes temporal sequences to capture dependencies over time. It is followed by an attention layer [14] for temporal focus, and fully connected layers leading to a softmax layer for classification purposes. LeakyReLU activations and dropout are used in this branch for non-linearity and regularization, respectively.

IV. EXPERIMENT

There are many speech emotion datasets. In this paper, we select EMODB and RAVDESS datasets for our experiments since they both contain noise-free emotional audio. In this way, we can well control the testing conditions in the experiments to investigate the robustness of SER systems clearly.

A. EMODB Dataset

EMODB is a German emotional speech dataset [42]. Ten actors (five males and five females) participated in the data recording. Each of them spoke ten sentences with different emotions. The audios were recorded with the 16 kHz sampling rate. After the quality evaluation, 535 utterances were included in this dataset. The dataset comprises seven emotions: anger,

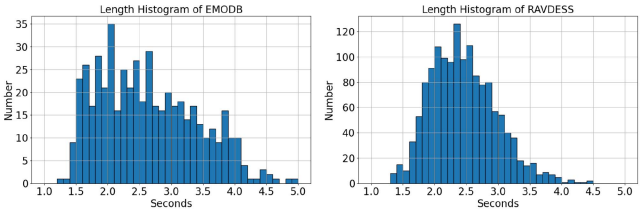

Fig. 5. Length distributions of audio signals from EMODB and RAVDESS datasets.

boredom, anxiety, happiness, sadness, disgust, and neutral. For our experiments, we included all 535 emotional utterances. We set the maximum sentence length to 3 seconds, cutting off longer sentences and zero-padding shorter sentences. The average length of audio signals in EMODB is 2.77 seconds and the histogram of audio lengths is shown in the left panel of Fig. 5.

B. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [43] features recordings from 24 actors (12 males and 12 females). Each actor delivers lexically matched statements in a neutral North American accent, expressing various emotions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. Each emotion, except for neutral, is expressed at two levels of intensity (normal and strong). The original audio sampling rate is 48 kHz, and we resampled audio clips to 16 kHz in this paper. Each actor interprets two phrases, "Kids are talking by the door" and "Dogs are sitting by the door," using eight emotions, and each emotional phrase is spoken twice.

Thus, there are a total of 1440 emotional phrases ($= 7$ (emotions, excluding neutral) $\times$ 2 (intensity levels) $\times$ 2 (phrases) $\times$ 2 (repetitions) $\times$ 24 (actors) $+$ 1 (neutral emotion) $\times$ 1 (intensity level) $\times$ 2 (phrases) $\times$ 2 (repetitions) $\times$ 24 (actors)). We included all 1440 emotional phrases for our experiments. We also set the maximum sentence length to 3 seconds, cutting off longer sentences and zero-padding shorter sentences. The average lengths of audio signals in RAVDESS is 2.45 seconds and the histogram of audio lengths is shown in the right panel of Fig. 5.

C. Experiment Setups

In this paper, the Mel-spectrogram was obtained using following parameters: a window length of 40 milliseconds, a hop size of 10 milliseconds, a 2048-point FFT, and 128 Mel-frequency bins within the range of 62.5 Hz to 8000 Hz. For generating the spectro-temporal modulation features, the rate ω_0 was set to ($\pm$2, $\pm$4, $\pm$8, $\pm$16, $\pm$32) Hz, and the scale Ω to (0.5, 1, 2, 4) cycles/20 Mel-bands. Consequently, the size of the spectro-temporal modulation feature of an audio clip is 40*T*128, where T is the number of time frames.

We conducted 10-fold cross-validation on both datasets. Each dataset was randomly divided into ten portions, with nine portions allocated for training the model and one for validation. That is, we evaluated the model under the speaker-dependent condition as in [44], [45]. This procedure was repeated ten times, with each portion used as validation data once, and the performance for each fold averaged. The model was trained using the Adam optimizer with cross-entropy loss, employing a dropout rate of 0.1, a learning rate of 0.001, and a batch size of 64. Additionally, label smoothing with a factor of 0.1 [46] was implemented for regularization. Each dataset underwent 10-fold cross-validation five times, and we reported the five runs' mean and standard deviation.

D. Evaluation Metric

In this study, we assessed model performance using two metrics. First, we employed the unweighted average recall (UAR), defined as follows:

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$UAR = \frac{1}{N}\sum_{i=1}^{N} Recall_i \tag{7}$$

where i is the index of class, TP_i is true positives of i-th class (i.e., the number of correct predictions in i-th class), and FN_i denotes false negatives of the i-th class (i.e., the number of incorrect predictions in the i-th class). N represents the total number of classes. Second, we utilized the weighted average recall (WAR), also referred to as accuracy, defined as:

$$\text{WAR} = \frac{\text{total number of correct predictions}}{\text{total number of predictions}} \tag{8}$$

These metrics provided comprehensive insights into the model's performance across different classes.

V. Results and Discussion

A. Clean Condition

Tables IV shows scores of UAR and WAR of the compared models under noise-free conditions using EMODB and RAVDESS datasets, respectively. The tACRNN represents the proposed two-stream model with the modulation branch, the spectrogram branch, and the recognition branch. The ACRNN_STM represents the model with the modulation branch and the recognition branch, and the ACRNN_Mel represents the model with the spectrogram branch and the recognition branch. 3D-ACRNN [14], TIM-Net [44] and MS-SENet [45] are the compared models retrained on our dataset using the codes from the GitHub repository. As shown in the table, ACRNN_STM using spectro-temporal modulation features performs worse than the ACRNN_Mel with log Mel-spectrogram, TIM-Net, MS-SENet and 3D-ACRNN when evaluated using EMODB dataset. However, simultaneously using both the log Mel-spectrogram and modulation features boosts the performance to approach that of the compared models. Similar results can be observed on RAVDESS dataset.

These results suggest that MFCC and log Mel-spectrogram contain rich information for SER under clean conditions. On the other hand, the spectro-temporal modulation features, generated by filtering the log Mel-spectrogram through a limited number of spectro-temporal modulation filters, contain less information than the original log Mel-spectrogram, such that ACRNN_STM

TABLE IV
MODEL PERFORMANCE (%) FOR SPEAKER-DEPENDENT SER USING EMODB AND RAVDESS DATASETS

Model	Feature	Param.	EMODB		RAVDESS	
			UAR	WAR	UAR	WAR
[47]*	MFCC+GTCC	-	94.30	94.58	88.54	89.10
[48]*	STFT SPEC	-	93	91	89	82
TIM-Net [44]	MFCC	104K	93.26±0.62	93.34±0.49	91.94±0.37	91.76±0.34
MS-SENet [45]	MFCC	122K	93.40±0.35	93.79±0.28	89.60±0.20	89.50±0.15
3D-ACRNN [14]	log Mel-SPEC+Δ+ΔΔ	3.4M	91.88±0.72	92.60±0.4	88.74±0.12	88.81±0.32
ACRNN_Mel	log Mel-SPEC	7.72M	92.76±0.77	93.27±0.60	87.24±0.69	87.26±0.57
ACRNN_STM	STM	7.79M	86.05±0.63	87.06±0.32	79.28±1.12	79.65±0.80
tACRNN	log Mel-SPEC + STM	15.37M	90.87±0.49	91.47±0.45	83.55±0.80	83.33±0.39

GTCC is short for Gamma-tone cepstral coefficient, STFT is short for short-term fourier transform, SPEC is short for spectrogram, STM is short for spectro-temporal modulation, and Δ represents the delta feature. 3D-ACRNN uses 40 Mel-bands while ACRNN_Mel and tACRNN use 128 Mel-bands to match the setting of the auditory model. Note, the scores of the asterisked models (*) are retrieved from their original papers.

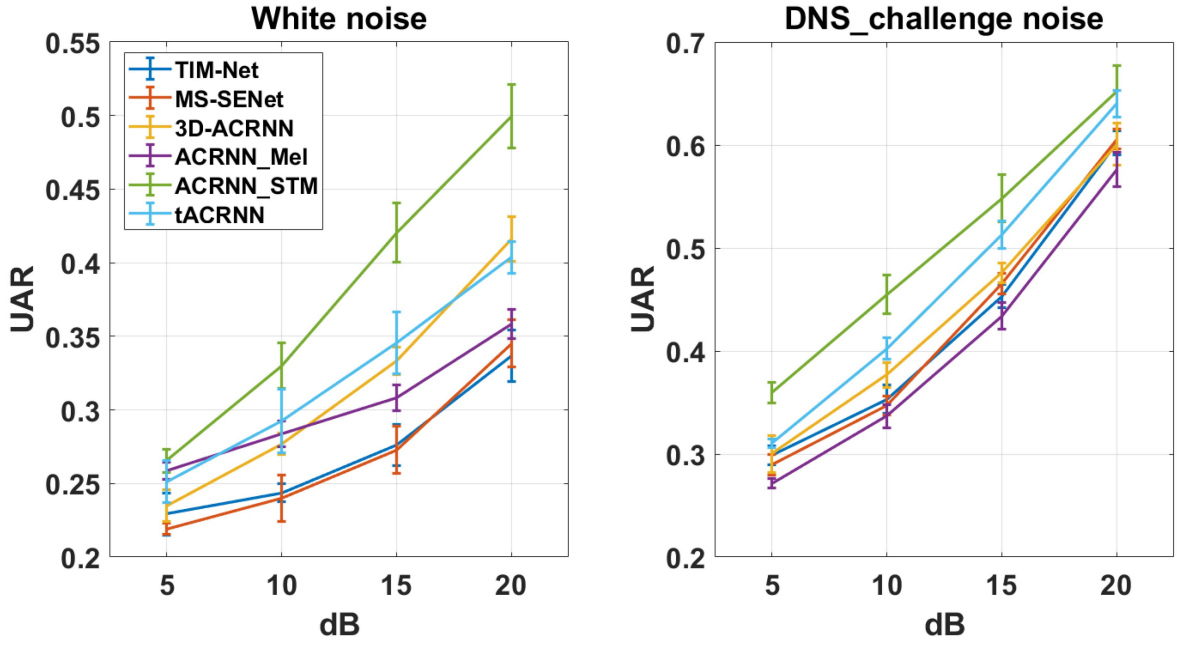

Fig. 6. Robustness test scores with white and DNS challenge noise on EMODB dataset.

performs worse than ACRNN_Mel. In the proposed tACRNN model, although features from both the spectrogram and modulation branches are concatenated, the modulation features tend to degrade the performance of the spectrogram features under the clean condition. This degradation might be alleviated by incorporating a noise condition-aware feature fusion method to replace the direct concatenation.

B. Noise Robustness

In this section, we demonstrate the performance of SER models when exposed to additive noise. Specifically, we conducted clean-train-noisy-test experiments. To create noisy evaluation samples, we used samples in EMODB and RAVDESS as clean sources. Each sample was then added with different noise extracted from the DNS challenge [49] and white noise at signal-to-noise ratios (SNRs) of 5, 10, 15, and 20 dB. The noisy clips were not seen during the training phase.

Fig. 6 illustrates the performance of the proposed model and the compared models in terms of UAR scores with different SNR conditions on EMODB. We report the mean and standard deviation of UAR from five runs of 10-fold cross-validation. Based on the mean values shown in the left panel of the figure, under white noise, ACRNN_STM and tACRNN surpass ACRNN_Mel at 10, 15, and 20 dB, and consistently outperform TIM-Net and MS-SENet at all SNR levels. As for comparing to 3D-ACRNN, ACRNN_STM and tACRNN respectively achieve higher performance at all SNR levels, and at 5, 10, and 15 dB. In addition to comparing mean values, statistical analysis was further performed using a pair-wise t-test. A total of 36 pairs of t-test were conducted, including tACRNN versus each of the other five models at each SNR, and ACRNN_STM versus each of the four models, ACRNN_Mel, TIM-Net, MS-SENet, and 3D-ACRNN, at each SNR. Out of the 36 pairs of t-test, 27 tests show the differences of performance of compared models are statistically significant (p-value < 0.05). The other 9

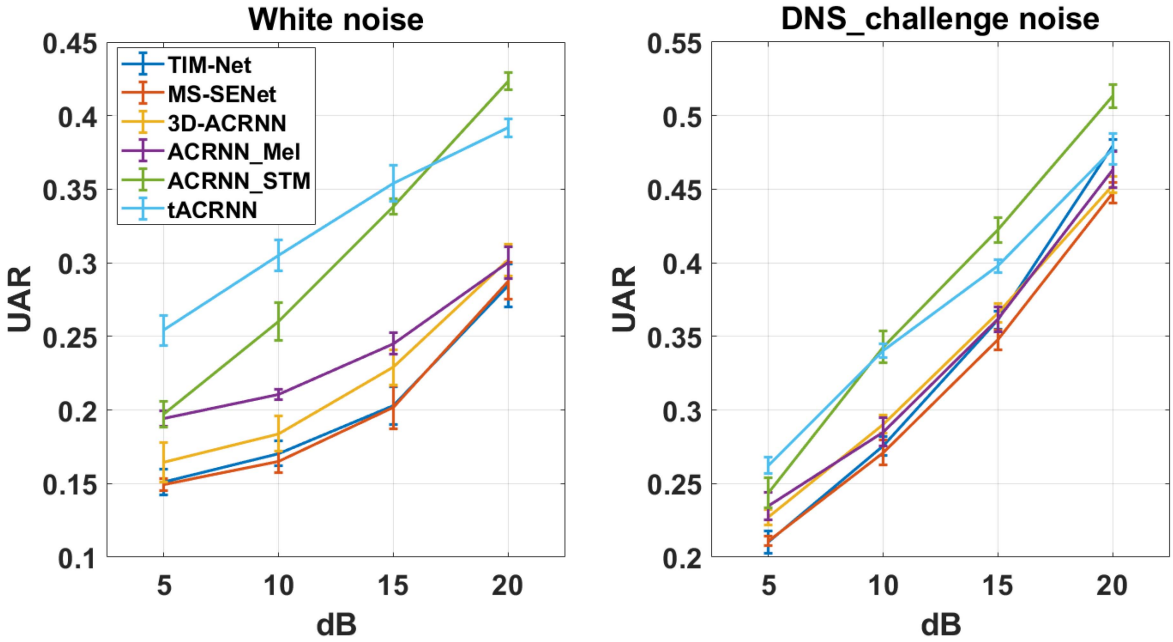

Fig. 7. Robustness test scores with white and factory noise on RAVDESS dataset.

pairs with no significant difference include <tACRNN versus ACRNN_STM > at 5 dB, < tACRNN versus ACRNN_Mel > at 5 and 10 dB, < tACRNN versus 3D-ACRNN > at all SNR levels, < tACRNN versus TIM-Net > at 5 dB, and < ACRNN_STM versus ACRNN_Mel > at 5 dB. Under DNS challenge noise, the mean values in the right panel show ACRNN-STM and tACRNN consistently outperform TIM-Net, MS-SENet, 3D-ACRNN, and ACRNN-Mel at all SNR levels. Meanwhile, statistical analysis shows only 3 out of 36 tests produce no significant difference. These 3 t-test conditions are < tACRNN versus ACRNN_STM > at 20 dB, < tACRNN versus TIM-Net > at 5 dB, and < tACRNN versus 3D-ACRNN> at 5 dB.

Fig. 7 illustrates the performance of the compared models on the RAVDESS dataset under different SNR conditions. As demonstrated from the left panel, both tACRNN and ACRNN_STM outperform ACRNN-Mel, TIM-Net, MS-SENet, and 3D-ACRNN under white noise. As for statistical analysis, only 1 out of 36 tests, i.e., <ACRNN_STM versus ACRNN_Mel > at 5 dB, produces no significant difference. On DNS challenge noise, results in the right panel show both tACRNN and ACRNN_STM almost consistently outperform the other four models over tested SNRs. However, statistical analysis indicates 4 out of 36 tests produce no significant difference. These 4 t-test conditions are < tACRNN versus ACRNN_STM > at 10 dB, < tACRNN versus ACRNN_Mel > and < tACRNN versus TIM-Net > at 20 dB, and < ACRNN_STM versus ACRNN_Mel> at 5 dB.

Results in Figs. 6 and 7 demonstrate models incorporating modulation features, namely ACRNN_STM and tACRNN, mostly exhibit higher performance than ACRNN_Mel, TIM_Net, MS-SENet and 3D-ACRNN at 10, 15, and 20 dB SNR conditions. Conversely, although TIM_Net and MS-SENet perform better in clean conditions (as shown in Table IV), they deteriorate significantly in the presence of noise. Interestingly, results show 3D-ACRNN performs comparably to the proposed tACRNN on EMODB under white noise. In Section V-C1, we will speculate the causes of 3D-ACRNN's robustness which only appears under this particular condition.

C. Rate-Scale Domain Analysis

1) Modulation Energy Distribution: To show the spectro-temporal modulation features are indeed less noise-degraded features, we derive the modulation energy distribution in the rate-scale domain by averaging spectro-temporal modulation features across time and frequency axes as follows:

$$E\left(\pm\omega_0, \Omega_0\right) = \sum_n \sum_k \mid C\left(n, k; \pm\omega_0, \Omega_0\right) \mid \qquad (9)$$

Fig. 8 displays the average modulation energy distributions of white noise, noise in DNS challenge, and clean speech samples from the EMODB and RAVDESS datasets. For display purposes, linear interpolation is done before plotting figures. The analysis reveals that the modulation power of white noise is mostly concentrated around the fast rate ($\omega = \pm 16, \pm 32$) and fine scale ($\Omega = 4$) region, while the modulation power of speech is mostly concentrated around the slow rate ($\omega = \pm 4, \pm 8$) and coarse scale ($\Omega < 2$) region. This figure suggests that noise and speech are quasi-separable in the rate-scale domain. Hence, we can infer that spectro-temporal modulation features are robust against noise.

In addition, based on results in Figs. 6 and 7, we observe the benefit of noise robustness of tACRNN against ACRNN_Mel is more prominent on the RAVDESS dataset than on the EMODB

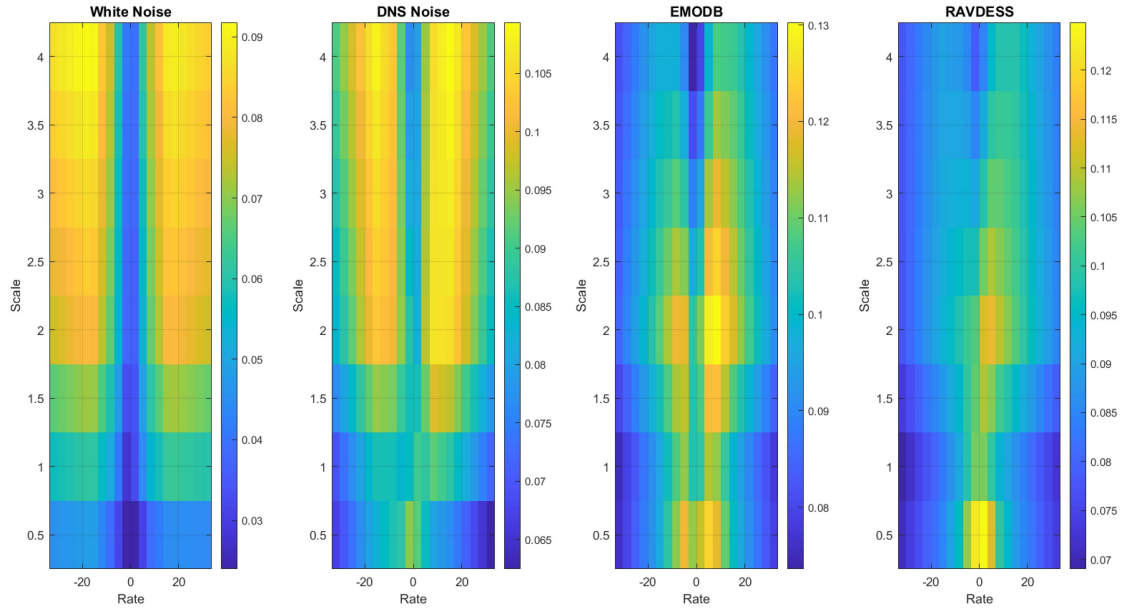

Fig. 8. Spectro-temporal modulation energy distributions of white noise, DNS challenge noise, clean speech clips in EMODB and RAVDESS dataset.

dataset under white noise. The right two panels in Fig. 8 reveal that STM features of RAVDESS speech concentrate at lower rates compared to the features of EMODB speech, probably due to the language differences between these two datasets. On the other hand, the STM features of white noise are located at higher rates than those of DNS challenge noise. Consequently, in the rate-scale domain, there is less overlap between white noise and the RAVDESS dataset compared to the overlap with the EMODB dataset. However, the difference in overlaps between both datasets to DNS challenge noise is relatively small. This explains why the improvement of robustness from ACRNN_Mel to tACRNN is most pronounced on the RAVDESS dataset under white noise.

Another interesting observation from Figs. 6 and 7 is that 3D-ACRNN exhibits certain robustness to white noise on the EMODB dataset. Such robustness is not observed either on the RAVDESS dataset or under DNS challenge noise. We speculate that this robustness comes from the use of the delta-delta feature in 3D-ACRNN. Based on the equations in [14], its delta feature is calculated using two preceding and two following time frames with a 10 ms hop size and a 25 ms window length. Therefore, its delta-delta feature, with a receptive field of approximately 105 ms, can capture temporal variations around 10 Hz. As shown in Fig. 8, EMODB speech contains much more components at rates around 10 Hz compared to RAVDESS speech. The proper receptive field combined with characteristics of the delta feature likely contribute to the effectiveness of 3D-ACRNN against stationary noise, such as white noise, on the EMODB dataset.

2) Rate-Scale Weights for Robust SER: To better understand the benefits induced by the modulation branch, we examine the weights of the first convolutional layer in the modulation branch of the well-trained tACRNN and ACRNN_STM models to discern the significance of rate-scale coded modulations on robust SER. First, we sum the weights of the convolutional layer across height, width, and output channels as follows:

$$\hat{W}(c_{in}) = \sum_{c_{out}=1}^{128} \sum_{m=1}^{5} \sum_{n=1}^{3} \mid W(c_{out}, c_{in}, m, n) \mid \tag{10}$$

where W contains the weights of 2D convolution kernels of the first convolutional layer in the modulation branch. c_{out}, c_{in}, m, and n are the indexes of the output channel, the input channel, height, and width, respectively. Second, we normalize the weight values to 0 to 1 as follows:

$$V_{\max} = \max_{c_{in}} \hat{W}(c_{in}) \tag{11}$$

$$V_{\min} = \min_{c_{in}} \hat{W}(c_{in}) \tag{12}$$

$$\hat{W}_{\text{norm}} = \frac{\hat{W} - V_{\min}}{V_{\max} - V_{\min}} \tag{13}$$

We observed that modulation branches of tACRNN and ACRNN_STM generate similar normalized weight patterns $\hat{W}_{\text{norm}}$. Thus, we only focus on the weight pattern of tACRNN in this paper.

The left panel of Fig. 9 shows the averaged $\hat{W}_{\text{norm}}$ of tACRNN trained on the EMODB dataset across 10 folds and 5 runs. The weight pattern indicates that tACRNN prioritizes the outputs of the modulation filters tuned to (rate, scale) = (+2, 4) and (−2, 1). The impulse responses of the two modulation filters are plotted in Figs. 10 and 11, with the vertical axis showing the Mel-band index and the corresponding frequency. The modulation filter, tuned to the scale of 4 cycles per 20 Mel-bands, captures spectral repetitive patterns of the period of 5 Mel-bands along the frequency axis. Within the range of the 1st Mel-band to the 20th Mel-band, the averaged frequency span of 5 Mel-bands is around 89.08 Hz, while within the range of the 81th Mel-band

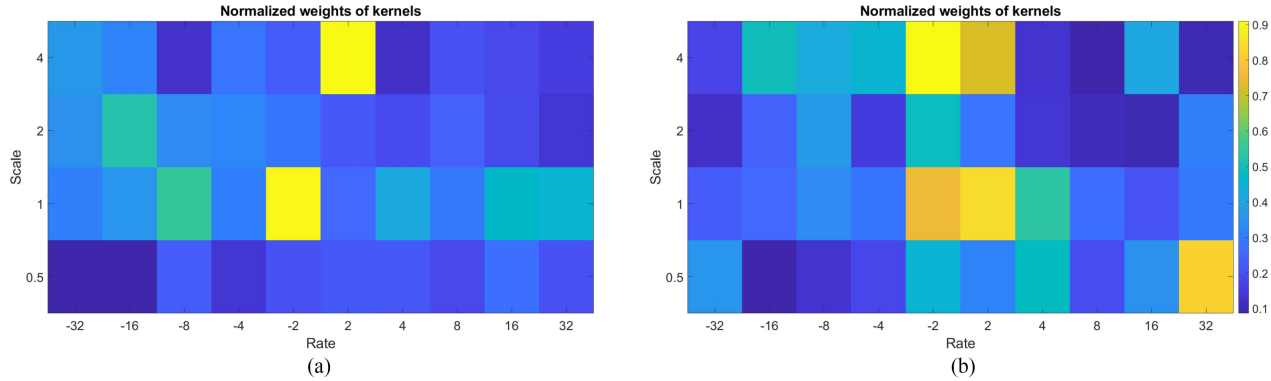

Fig. 9. Normalized weights of kernels of the first convolutional layer in the modulation branch. The weights are trained using the (a) Left pannel: EMODB (b) Right pannel: RAVDESS dataset and plotted against rate and scale parameters.

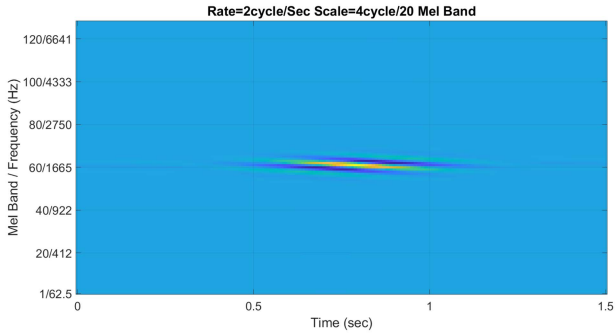

Fig. 10. Impulse response of the spectro-temporal modulation filter tuned to $(\omega, \Omega) = (+2, 4)$.

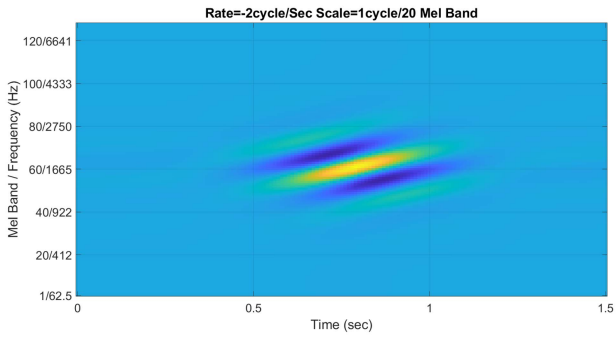

Fig. 11. Impulse response of the spectro-temporal modulation filter tuned to $(\omega, \Omega) = (-2, 1)$.

to the 100th Mel-band, the averaged frequency span of 5 Mel-bands becomes around 403.12 Hz. In other words, below the 100th Mel-band ($\simeq$4300 Hz), this modulation filter captures the spectral repetitive patterns with periods ranging from 89 Hz to 403 Hz. It suggests that this modulation filter focus on analyzing the harmonic structure of speech. Additionally, the downward direction emphasizes prosodic elements with declining patterns, aligning with findings by Rodero [50] on the importance of pitch contour in perceiving speech emotion. The other modulation filter, tuned to the scale of 1 cycle per 20 Mel-bands, captures spectral repetitive patterns of the period of 20 Mel-bands along the frequency axis. That is, below 4300 Hz, this modulation filter captures the spectral repetitive patterns with periods ranging from 335 Hz to 1516 Hz. It suggests this modulation filter focus on analyzing the formant structure of speech. As for emphasizing the rate of 2 Hz, we postulate the model pays more attention to long term ($\simeq$ 500ms) rather than short term temporal information for robust SER.

The right panel of Fig. 9 shows the average $\hat{W}_{\text{norm}}$ of tACRNN trained on the RAVDESS dataset across 10 folds and 5 runs. Similar to the weight pattern shown in the left pannel, the proposed two-stream model also prioritizes the outputs of the modulation filters tuned to (rate, scale) = (+2, 4) and (−2, 1), which analyze the long-term harmonic and formant contours. In addition, the model also put strong weights on filters tuned to (rate, scale) = (−2, 4) and (+2, 1). It suggests the model trained on the RAVDESS dataset focuses on both the upward and downward harmonic and formant contours. It is probably due to the language differences between German and North American English. In addition to long-term harmonic and formant contours, the model also pays attention to a type of broadband transient feature, which is resolved by the modulation filter tuned to (rate, scale) = (+32, 0.5). This phenomenon is not observed in the model trained on EMODB. We believe the model focusing on a broadband transient feature is an artifact caused by data preparation. As shown from histograms in Fig. 5, many more samples from RAVDESS are zero-padded than from EMODB, such that the model pays more attention to the artificial pattern of speech offset on samples from RAVDESS than from EMODB.

VI. Implications and Limitations

The broader implications of this work suggest that integrating STM features into the neural network model can significantly improve its robustness on SER in challenging acoustic environments, potentially influencing future research and development in this field. While this study demonstrates promising results, it is important to acknowledge several limitations. First, the model was tested under the speaker-dependent condition and has not yet been evaluated in the speaker-independent scenario. Second, we have primarily combined the STM features with log-Mel

spectrograms, but the integration of STM features with other commonly used representations on SER has not been explored.

VII. Conclusion

In this paper, we investigate the efficacy of incorporating spectro-temporal modulation features for robust SER. Our findings demonstrate that the model using only spectro-temporal modulation features performs worse than the model using acoustic features, Mel-spectrogram and MFCC, under noise-free conditions. Using the clean-train-noisy-test paradigm, the model integrating modulation features with log Mel-spectrogram features exhibit superior performance to the compared baseline models across various signal-to-noise ratios (SNRs) and noise types, including white and DNS challenge noise.

Our analysis reveal that the proposed two-stream model emphasizes specific spectro-temporal modulations related to the harmonic and formant contours on the Mel-spectrogram. This finding provides valuable insights into the mechanisms underlying SER. Moving forward, we aim to refine our approach by optimizing the rate-scale parameter selection process to prioritize features most relevant for emotion recognition. Additionally, we plan to extend our investigation to encompass reverberation to assess the robustness and generalizability of the proposed model.

References

[1] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. 2003 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2003, pp. II–1.

[2] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *Proc. 5th Int. Conf. Knowl. Smart Technol.*, 2013, pp. 86–91.

[3] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 223–227.

[4] A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition," in *Proc. 2019 IEEE Int. Conf. Signal Process. Inf. Commun. Syst.*, 2019, pp. 122–125.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. 2017 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 2227–2231.

[6] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proc. 2019 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 7405–7409.

[7] A. Dhall, O. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *Proc. 2015 ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, 2015, pp. 423–426. [Online]. Available: https://doi.org/10.1145/2818346.2829994

[8] F. Zhu-Zhou, D. Tejera-Berengué, M. Rosa-Zurera, M. Utrilla-Manso, and R. Gil-Pita, "Robust energy-efficient audio-based anger detection system for noisy environments," in *Proc. IEEE 19th Int. Conf. Intell. Comput. Commun. Process.*, 2023, pp. 405–411.

[9] P. Heracleous, K. Yasuda, F. Sugaya, A. Yoneyama, and M. Hashimoto, "Speech emotion recognition in noisy and reverberant environments," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 262–266.

[10] A. A. Abdelhamid et al., "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022.

[11] S. G. Leem, D. Fulford, J. P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *Proc. 2022 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 6447–6451.

[12] S. R. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1892–1905, Fourth Quarter, 2022.

[13] T. S. Chi, L. Y. Yeh, and C. C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," *J. Ambient Intell. Humanized Comput.*, vol. 3, pp. 47–60, 2012.

[14] 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.

[15] F. G. Zeng et al., "Speech recognition with amplitude and frequency modulations," in *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 7, pp. 2293–2298, 2005. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0406460102

[16] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, Jun. 2013. [Online]. Available: https://doi.org/10.1007/s10772-012-9172-2

[17] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 680–688, Second Quarter, 2022.

[18] M. Bhaykar, J. Yadav, and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM," in *Proc. IEEE 2013 Nat. Conf. Commun.*, 2013, pp. 1–5.

[19] Y. Xia, L. W. Chen, A. Rudnicky, and R. M. Stern, "Temporal context in speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3370–3374.

[20] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. 2022 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 6922–6926.

[21] H. Tang, S. M. Chu, M. Hasegawa-Johnson, and T. S. Huang, "Emotion recognition from speech via boosted gaussian mixture models," in *Proc. 2009 IEEE Int. Conf. Multimedia Expo*, 2009, pp. 294–297.

[22] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2006, pp. 809–812.

[23] S. S. Poorna, C. Y. Jeevitha, S. J. Nair, S. Santhosh, and G. J. Nair, "Emotion recognition using multi-parameter speech feature classification," in *Proc. 2015 Int. Conf. Comput. Commun. Syst.*, 2015, pp. 217–222.

[24] J. W. Mao, Y. He, and Z. T. Liu, "Speech emotion recognition based on linear discriminant analysis and support vector machine decision tree," in *Proc. 37th Chin. Control Conf.*, 2018, pp. 5529–5533.

[25] X. Yuan, W. P. Wong, and C. T. Lam, "Speech emotion recognition using multi-layer perceptron classifier," in *Proc. IEEE 10th Int. Conf. Inf. Commun. Netw.*, 2022, pp. 644–648.

[26] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in *Proc. 2021 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3020–3024.

[27] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proc. 2019 IEEE Int. Conf. Signals Syst.*, 2019, pp. 40–44.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, I. Guyon et al. Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[29] C. Lu, H. Lian, W. Zheng, Y. Zong, Y. Zhao, and S. Li, "Learning local to global feature aggregation for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 1908–1912.

[30] A. R. Avila, M. J. Alam, D. O'Shaughnessy, and T. H. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3663–3667.

[31] R. Chakraborty, A. Panda, M. Pandharipande, S. Joshi, and S. K. Kopparapu, "Front-end feature compensation and denoising for noise robust speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3257–3261.

[32] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1691–1695.

[33] F. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.

[34] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2327–2331.
[35] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639310001470
[36] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 177–188, First Quarter, 2021.
[37] B. M. Calhoun and C. E. Schreiner, "Spectral envelope coding in cat primary auditory cortex: Linear and non-linear effects of stimulus characteristics," *Eur. J. Neurosci.*, vol. 10, no. 3, pp. 926–940, 1998. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1460-9568.1998.00102.x
[38] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," *J. Neurophysiol.*, vol. 76, no. 5, pp. 3503–3523, 1996. [Online]. Available: https://doi.org/10.1152/jn.1996.76.5.3503
[39] T. S. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoustical Soc. America*, vol. 118, no. 2, pp. 887–906, Aug. 2005. [Online]. Available: https://doi.org/10.1121/1.1945807
[40] T. S. Chi and C. C. Hsu, "Multiband analysis and synthesis of spectro-temporal modulations of Fourier spectrogram," *J. Acoustical Soc. America*, vol. 129, no. 5, pp. EL190–EL196, Apr. 2011. [Online]. Available: https://doi.org/10.1121/1.3565471
[41] K. M. Cheong, Y. L. Shen, and T. S. Chi, "Active acoustic scene monitoring through spectro-temporal modulation filtering for intruder detection," *J. Acoustical Soc. America*, vol. 151, no. 4, pp. 2444–2452, Apr. 2022. [Online]. Available: https://doi.org/10.1121/10.0010070
[42] F. Burkhardt et al., "A database of german emotional speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 1517–1520.
[43] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, pp. 1–35, May 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391
[44] J. Ye, X. C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
[45] M. Li, Y. Zheng, D. Li, Y. Wu, Y. Wang, and H. Fei, "MS-SENet: Enhancing speech emotion recognition through multi-scale feature fusion with squeeze-and-excitation blocks," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 12271–12275.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[47] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection," *IEEE Access*, vol. 9, pp. 122855–122871, 2021.
[48] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[49] H. Dubey et al., "ICASSP 2023 deep noise suppression challenge," *IEEE Open J. Signal Process.*, vol. 5, pp. 725–737, Mar. 2024.
[50] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *J. Voice*, vol. 25, no. 1, pp. e25–e34, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199710000378

Yih-Liang Shen received the BS degrees in electronics and electrical engineering from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2017. He is currently working toward the PhD degree with the Institute of Communication Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. His research interests include auditory model, speech processing, deep learning, and emotion recognition.

Pei-Chin Hsieh received the BS degree in electronics and electrical engineering from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2023. She is currently working toward the PhD degree with the Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. Her research interests include auditory modeling, speech processing, automatic singing evaluation, and deep learning.

Tai-Shih Chi received the PhD degree in electrical engineering from the University of Maryland, College Park, in 2003. From 2003 to 2005, he was a research associate with the University of Maryland. He joined the Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan, in 2005. His research interests include neuromorphic auditory modeling, soft computing, and speech analysis and processing.