PAPER

Enhancing the Robustness of Speech Anti-spoofing Countermeasures through Joint Optimization and Transfer Learning

Yikang WANG[†a)], Xingming WANG[††], Chee Siang LEOW[†], Qishan ZHANG[††], Ming LI[††b)], *Nonmembers*,
and Hiromitsu NISHIZAKI[†c)], *Senior Member*

SUMMARY

Currently, research in deepfake speech detection focuses on the generalization of detection systems towards different spoofing methods, mainly for noise-free clean speech. However, the performance of speech anti-spoofing countermeasure (CM) systems often does not work well in more complicated scenarios, such as those involving noise and reverberation. To address the problem of enhancing the robustness of CM systems, we propose a transfer learning-based hybrid approach with Speech Enhancement front-end and CounterMeasure back-end Joint optimization (SECM-Joint), investigating its effectiveness in improving robustness against noise and reverberation. Experimental results show that our SECM-Joint method reduces EER by 19.11% to 64.05% relatively in most noisy conditions and 23.23% to 30.67% relatively in reverberant environments compared to a Conformer-based CM baseline system without pre-training. Additionally, our dual-path U-Net (DUMENet) further enhances the robustness for real-world applications. These results demonstrate that the proposed method effectively enhances the robustness of CM systems in noisy and reverberant conditions. Codes and experimental data supporting this work are publicly available at: `https://github.com/ikou-austin/SECM-Joint`[†]

key words: *speech anti-spoofing, transfer learning, joint optimization*

1. Introduction

The emerging Deepfake Speech Detection (DSD) field aims to develop anti-spoofing countermeasure (CM) systems to detect synthesized speech, addressing risks to social security, political stability, and economic integrity [1]. As shown in Figure 1, a typical CM system pipeline consists of preprocessing, feature extraction and classification. If a deep neural network is employed to obtain learnable features directly from raw speech waveform input, replacing the separate feature extraction and classification processes, the entire CM system can be regarded as an end-to-end model [1].

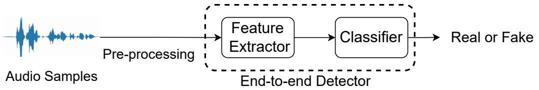


Fig. 1 The illustration of typtical pipeline solution for CM systems. For end-to-end models, feature extraction and classifiers are integrated into one neural network.

In DSD studies based on clean data, various approaches have been explored, including conventional Gaussian Mixture Model (GMM) [2] and neural network-based (NN-based) models, such as Light Convolutional Neural Network (LCNN) [3], Deep Residual Network (ResNet) [4], and Graph Neural Networks (GNNs) [5], [6]. GMM has been widely adopted as a baseline model in a series of competitions, such as ASVspoof 2017 and 2019 [7], [8]. However, most NN-based models outperform the GMM-based classifier [1]. In addition, various novel training strategies have been proposed to further enhance the performance of the CM system. For instance, Xue et al. [9] proposed a self-distillation network, where the deeper network guides shallower networks to better capture fine-grained information. Wang et al. [10] introduced the adaptive and hyper-parameter-free probability-to-similarity gradient (P2SGrad) into the DSD task. Moreover, several studies have explored end-to-end approaches as alternatives to separately designed feature extractors and classifiers. Among them, the spectro-temporal graph attention network (RawGAT-ST) [14], Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks (AASIST) [15], and the Squeeze-and-Excitation Rawformer (SE-Rawformer) [16] have shown promising results.

Although most research have primarily focused on DSD under clean-speech conditions [3]–[5], [7]–[16], some studies have focused on enhancing the robustness of DSD tasks in complex scenarios. For example, the ASVspoof 2021 Challenge logical access (LA) [17] was designed to investigate DSD under audio channel coding and compression conditions. Research works [18]–[21] on the robustness of CM system primarily focuses on the preprocessing and feature extraction stages of the pipeline. In the preprocessing stage, as illustrated in Figure 2(a), data augmentation is the most commonly employed technique to enhance the model's robustness. For instance, Tak et al. [18] proposed the RawBoost data augmentation method, which is specifically designed for telephony scenarios and does not require additional data sources, thereby effectively improving the robustness of CM systems against compression artifacts. Wang et al. [19] introduced techniques such as band trimming, band-pass filtering, and band extension to determine optimal sub-band widths for coding and compression robust

[†] Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Kofu, Japan

[††] Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, China

a) E-mail: g21dtsa1@yamanashi.ac.jp
b) E-mail: ming.li369@dukekunshan.edu.cn
c) E-mail: hnishi@yamanashi.ac.jp

DOI: 10.1587/transinf.E0.D.1

[†] The training code and detailed implementation will be released upon acceptance of this paper.

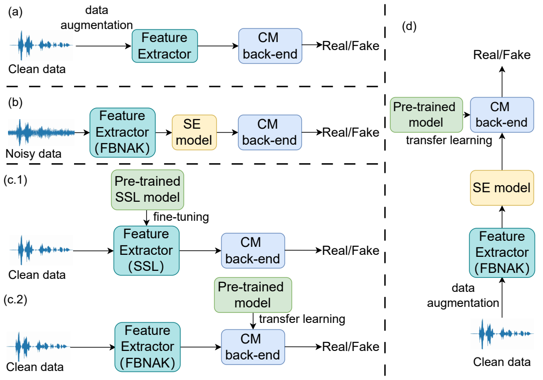


Fig. 2 (a) represents the training method using data augmentation, (b) represents the training method that integrates a speech enhancement front-end with joint optimisation of the fake speech detection system, (c.1) and (c.2) illustrates two different pre-training strategies, while (d) represents the SECM-Joint proposed in this work.

DSD tasks. In the feature extraction stage, Hanilçi et al. [20] compared different feature inputs under various noise conditions, proposed a feature fusion approach, and demonstrated that speech enhancement (SE) applied independently of CM training is ineffective for spoofing detection in noisy environments. Their findings demonstrate that environmental noise and reverberation can significantly degrade CM system performance. Additionally, Fan et al. [22] proposed a dual-branch knowledge distillation-based synthetic speech detection method and evaluated its performance under various noise conditions using a private dataset.

As shown in Figure 2(b), in our previous work [23], we introduced a joint optimization framework combining a SE front-end with an anti-spoofing back-end, demonstrating improved performance on noisy test data. However, this approach had several critical limitations. It focused solely on environmental noise without addressing reverberation effects. Additionally, the method required training separate models for each noise type and signal-to-noise ratio (SNR), while relying on offline-generated noisy datasets for each condition. Perhaps most importantly, the experimental setup limited assessment of model generalization under realistic conditions, constraining its practical applicability. To address these limitations, this paper makes several significant contributions to the field. We systematically evaluate CM robustness under both noise and reverberation conditions using on-the-fly data augmentation, enabling a single model to handle multiple acoustic conditions rather than requiring separate models for each scenario. We propose a Transfer Learning-based Speech Enhancement front-end and CounterMeasure back-end Joint optimization (SECM-Joint), as illustrated in Figure 2(d), which integrates transfer learning from pre-trained Automatic Speech Recognition (ASR) Conformer models with our joint optimization framework. This approach enhances system performance by leveraging prior knowledge from large-scale speech tasks. Furthermore, we introduce a Dual-input U-Net-based Masked Feature Enhancement Network (DUMENet), which uses masked features instead of direct speech reconstruction to more effectively handle non-additive acoustic distortions, particularly reverberation. Our comprehensive experiments across various noise types (environmental, babble, and music) at different SNRs (0-20dB) and reverberation conditions (RT60 from 0.25s to 1s) demonstrate that SECM-Joint significantly improves CM system robustness. Results show relative Equal Error Rate (EER) reductions of 19.11% to 64.05% in most noisy conditions and 23.23% to 30.67% in reverberant environments compared to baseline systems without pre-training. This work bridges an important gap between laboratory DSD research and practical applications by enhancing CM robustness in adverse acoustic environments, advancing more reliable spoofing detection for real-world deployment. The key contributions of this work are as follows:

- Refining the previously proposed joint optimization framework by evaluating SE model for spoof detection under noise and reverberation conditions.
- Proposing SECM-Joint, a transfer learning-based joint optimization approach that integrates a pre-trained Conformer model—originally trained on ASR tasks—into the DSD task, demonstrating that transfer learning from ASR tasks significantly enhances CM system robustness against noise and reverberation.
- Introducing DUMENet, a dual-input U-Net-based masked feature enhancement network, which leverages masked features instead of direct reconstruction FBANK to effectively handle the reverberation.

2. Methodology

To distinguish the authenticity of mixed speech signals with noise or reverberation, we summarize three training strategies that CM systems can use: data augmentation, speech enhancement, and utilizing pre-trained models, as shown in Figure 2(a), (b) and (c).

2.1 Data Augmentation in the Pre-processing

Data augmentation involves adding noise or reverberation to clean speech data, making the input speech more diverse. Models trained using data augmentation techniques have a stronger ability to handle complex speech signals. In this study, we use the officially provided ASVspoof 2019 LA (19LA) train subset as the clean data, adding noise or reverberation to obtain the corresponding mixed training data. According to [24], to obtain simulated reverberant signals, we usually convolve the clean source speech with the room impulse response (RIR). When using data augmentation methods, we perform on-the-fly noise and reverberation simulation on the input data, making the training samples more diverse. We ensure that the noise samples used for the training and test sets are isolated. For details on the implementation of online data augmentation and the generation of test datasets, see Chapter 3.

2.2 Speech Enhancement Module

The purpose of SE is to remove noise and reverberation from

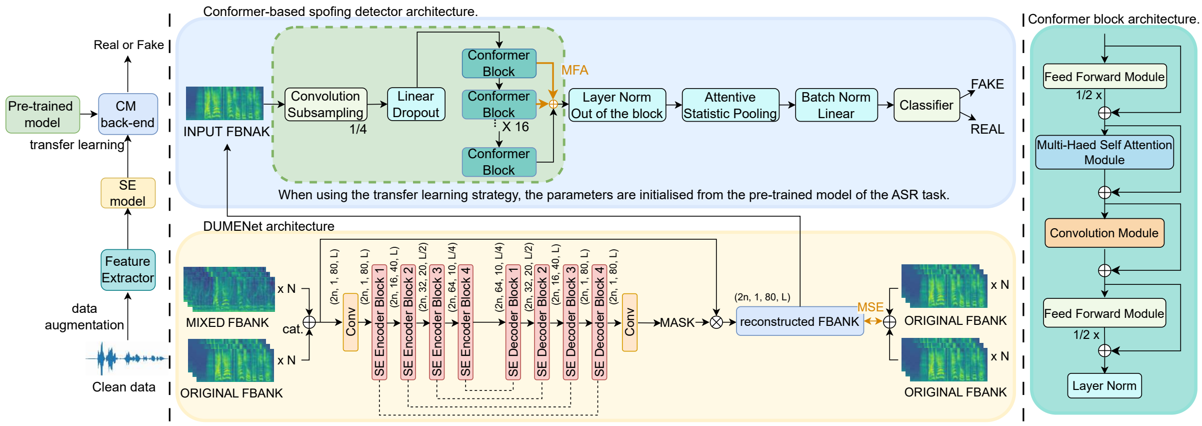


Fig. 3 Model architectrue of proposed SECM-Joint and DUMENet. For experiments without the SE module, the input feature is fed directly into the back-end spoofing detector.

the mixed speech signal $x(t)$ and estimate the target clean speech $s(t)$. Using the basic Unet model in the temporal-frequency domain as an example, an SE network applies the short-time Fourier transform (STFT) to the input waveform $x(t)$, decomposing the complex-valued spectrogram into magnitude and phase components. Only the magnitude is then fed into the Unet enhancement network, which returns an estimated magnitude spectrogram of the clean speech. To generate the corresponding clean speech waveform, the spectrogram is combined with the mixed phase and converted back to the time-domain waveform $s_i(t)$ via inverse STFT or some hand-crafted vocoder algorithms [22]. This process can be represented as follows:

$$|X(t,f)|e^{j\phi(t,f)} = \mathrm{STFT}(x(t)) \tag{1}$$

$$|S_i(t,f)| = \mathrm{Unet}(|X(t,f)|) \tag{2}$$

$$s_i(t) = \mathrm{iSTFT}(|S_i(t,f)|e^{j\phi(t,f)}) \tag{3}$$

where $Unet(\cdot)$ is the SE network, $|X(t,f)|$ is the magnitude spectrogram, and $\phi(t,f)$ is the phase spectrogram. For a pair of clean and mixed magnitude spectrograms $|S(t,f)|$ and $|X(t,f)|$, only the magnitude $|X(t,f)|$ is input into SE model, and the Unet model returns the estimated magnitude spectrogram $|S_i(t,f)|$, where f represents the frequency bin index and t represents the time frame. The loss function L_{mse} is designed to minimize the mean square error (MSE) between the clean spectrum and the recovered spectrum,

$$L_{mse} = \frac{1}{tf}\sum_t\sum_f \|\ |S_i(t,f)| - |S(t,f)|\ \|_2^2 \tag{4}$$

where $||\cdot||_2$ denotes the $L2$ norm.

In this paper, we propose DUMENet, as shown in Figure 3, whose model structure is almost the same as the Unet-based front-end SE network structure in our previous work [23]. Inspired by the work of Kim et al. [25] on the joint training of SE front-ends with speaker embedding extractors in the automatic speaker verification (ASV) task, we concatenates the log Mel-filterbank (FBANK) features of both the mixed speech and clean speech as the input at every training minibatch and uses the FBANK features of the clean speech, concatenated twice, as the targets labels. In contrast to using only mixed speech as input with clean labels, the dual-feature approach also optimizes clean speech inputs, ensuring denoising without introducing artifacts via redundant clean speech modifications [25]. Per the Convolution Property [26], time-domain convolution corresponds to frequency-domain multiplication: $\mathcal{F}\{x(t) * h(t)\} = X(j\omega)H(j\omega)$, where $x(t)$ and $h(t)$ are time-domain signals, $*$ denotes convolution, and $\mathcal{F}\{x(t)\} = X(j\omega)$ and $\mathcal{F}\{h(t)\} = H(j\omega)$ represent their Fourier transforms.

Unlike our previous work [23], DUMENet does not directly reconstruct the FBANK features of the clean speech. Instead, it outputs a feature shape mask. This mask is element-wise multiplied with the input feature, and the resulting product is compared to the label using the MSE loss function. This approach aligns with the reverberation simulation principles mentioned in Section 2.1, further enhancing the performance of the CM system under reverberant conditions. The loss function $L_{mse'}$ for DUMENet can be expressed as:

$$L'_{mse} = \frac{1}{Td}\sum_T\sum_d \|X(T,d)\mathbf{M}(T,d)) - S(T,d)\|_2^2 \tag{5}$$

where $X(T,d)$ and $S(T,d)$ represent the concatenated input and label FBANK features with a time series length of T and a Mel-scale filter bank dimension of d, respectively. $\mathbf{M}$ denote the output mask of DUMENet.

2.3 Transfer Learning in the Countermeasure Back-end

Unlike Tak et al., who used the large-scale self-supervised pre-trained model wav2vec2.0 as a front-end feature extractor for fine-tuning [13]. In this study, we use a pre-trained Conformer model [27] as the CM back-end. The Conformer model is trained on ASR tasks and has achieved excellent results in ASR tasks [28]. ASR pre-training enhances the generalisation of the CM system, complements front-end pre-training, and allows greater flexibility in front-end feature representations when applied in the back-end.

We employ the multi-scale feature aggregation-Conformer (MFA-Conformer) framework [29] as the CM back-end. As shown in Figure 3, this framework concatenates the output feature maps of all L Conformer blocks to form a large feature map $\mathbf{H}'$:

$$\mathbf{H}' = \mathrm{Concat}(h_1, h_2, \ldots, h_L) \tag{6}$$

where $\mathbf{H}'$ has dimensions $\mathbb{R}^{D\times T}$, with $D = d \times L$.

For the transfer learning process, we load the pre-trained Conformer [27] encoder parameters from the ASR task into the corresponding the MFA-Conformer structure. Similar to Cai et al. [30] leveraging the ASR-pretrained Conformer-MFA model for ASV tasks, we use MFA to concatenate the output feature maps from each Conformer block, apply layer normalization to obtain frame-level high-level representations, and input these representations into an attentive statistics pooling (ASP). layer and a fully connected (FC) layer to obtain utterance-level spoofing detection embeddings. Finally, a linear classifier is attached for fine-tuning. Similar to training a general DSD network, we fine-tune the Conformer-based detector using a binary cross-entropy loss function as given by:

$$L_{BCE} = \sum_i -(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)) \tag{7}$$

where $y_i \in \{0, 1\}$ represents the labels and p_i represents the classifier's probability output.

2.4 Joint Optimization of Speech Enhancement Front-end and Countermeasure Back-end

2.4.1 Audio Anti-spoofing Module

In order to explore whether SE co-optimization after changing the training strategy is consistent with the findings of our previous work [23], in addition to the Conformer-based beck-end mentioned in the previous section, we used the CM back-end models that were used in our previous studies: the LCNN [3] and ResNet18 [31]. The Max-Feature-Map (MFM) [32] operation based on the max-out activation function is a fundamental part of LCNN, which uses a Bi-LSTM [33] layer to aggregate corpus-level embeddings. For ResNet18, it is a lightweight version of ResNet.

2.4.2 Transfer Learning-based Joint Optimization

As shown in Figure 3, the SECM-Joint combines the three previously mentioned schemes by using on-the-fly data augmentation at the pre-processing stage to increase the diversity of input speech. The CM back-end network uses a pre-trained Conformer model based on the ASR task to improve generalization performance. In between, we incorporate DUMENet to jointly optimize with the back-end network during training. This aims to reduce unknown additional artifacts due to the inconsistency between independent SE and anti-spoofing tasks, making the entire CM system more conducive to the deep embedding of the DSD task. Similar to the joint optimization loss used in our previous work [23], SECM-Joint also uses MSE loss and CE loss as a combined loss function, as shown below:

$$L = L_{ce} + L'_{mse} \tag{8}$$

3. Datasets

The detailed statistics of the databases used in this work are outlined in Table 1. The ASVspoof database is a series of data from the ASVspoof challenges [7], [8], [17], [34]. Among them, the 19LA dataset is widely regarded as a clean, noise-free dataset, and the most widely used datasets by DSD researchers. For this work, we use 19LA datasets as the noise-free experimental dataset. Noisy and reverberant datasets are generated based on this dataset. The 19LA dataset consists mostly of clean data created using utterances from 107 speakers from the VCTK dataset [35]. These 107 speakers are partitioned into three speaker-disjoint sets for training, development, and evaluation. The spoofed utterances were generated using four TTS and two VC algorithms in the training and development sets, while 13 TTS/VC algorithms are used in the evaluation set, 4 of which are partially known and 7 of which are unknown for training and development [8].

3.1 On-the-Fly Data Augmentation During Training

Two different on-the-fly data simulation strategies were used for data augmentation of noise and reverberation condition during training to ensure diversity of training data.

1) Additive Noise Augmentation: We selected the MUSAN dataset [36] as our noise source, which contains approximately 60 hours of English real **speech**, 42 hours and 31 minutes of various **music**, and about 6 hours of various machine and **environmental noises**. We selected only a few hundred samples from each noise category as the source noise for training and the rest is used for simulating evaluation data. During the noise addition process, a noise category is first randomly selected with equal probability across three categories, and an instance from the chosen category is mixed with the clean speech signal at a specified SNR. The SNR is randomly chosen from a range of 0 to 20 dB. Following the noise-adding method in [23], we add environmental, music, and babble noise $n(t)$ to the clean speech $s(t)$ to obtain the mixed speech $x(t)$.

2) Convolutional Reverberation Augmentation: We can obtain reverberant speech $x_r(t)$ through the convolution operation $x_r(t) = s(t) * h(t)$, by convolving the clean speech $s(t)$ with the RIR $h(t)$. Luo et al. proposed an RIR simulation tool called FRAM-RIR [37], which efficiently performs online reverberation augmentation on clean audio. When using this tool, we modified two parameters: we set the rectangular room dimensions to $3m\times3m\times2.5m \sim 10m\times6m\times4m$,

Table 1 Dataset Statistics for Training and Development set from 19LA, and offline generated Evaluation Sets. N-eval refers to the noise evaluation set, including environmental noise, babble noise, and music noise, with each dataset divided into 5 subsets at SNRs of 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. R-eval refers to the reverberation evaluation set, containing 4 subsets with RT60 values of 0.25 s, 0.5 s, 0.75 s, and 1 s.

Dataset	**Genuine**	**Fake**	**Duration**
train	2,580	22,800	3.68 h
dev.	2,548	22,296	3.66 h
N-eval. (env.)	7,355 ×5	63,882 ×5	9.38 h ×5
N-eval. (babble)	7,355 ×5	63,882 ×5	9.38 h ×5
N-eval. (music)	7,355 ×5	63,882 ×5	9.38 h ×5
R-eval.	7,355 ×4	63,882 ×4	9.38 h ×4

and the reverberation time (RT60) was randomly chosen between 0.2 to 1.0 s. RT60 is the main room acoustic parameter, representing the time required for the sound energy in a room to decrease by 60 dB after the source emission has stopped. Generally, a larger RT60 indicates stronger reverberation.

To ensure training variability, we applied on-the-fly data augmentation with 0.7 probability per sample, resulting in 70% augmented and 30% unchanged data per batch.

3.2 Preparation of Evaluation Datasets

To distinguish the effects of noisy and reverberant data on DSD task, we created two evaluation datasets by adding noise and reverberation to the 19LA evaluation set offline.

1) Noise Evaluation Datasets: To create the evaluation sets, we randomly sampled from over 100,000 unseen noise of the MUSAN corpus, specifically utilizing the portion excluded from training data. The instances each category were mixed with clean speech at five SNR levels, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB, generating 15 evaluation subsets, as shown in Table 1. These 15 noise evaluation datasets are the same test data used in our previous study [23].

2) Reverberation Evaluation Datasets: Although the study by Tom et al. provides about 40,000 simulated RIRs [38], they roughly categorize rooms into large, medium, and small sizes. To compare the effects under different reverberation conditions, we use RT60 as an indicator of reverberation intensity. When creating reverberation test datasets, we used the pyroomacoustics[†] toolkit to generate RIRs offline. During the simulation, we randomly set the room dimensions to $10m \times 8m \times 2.8m \sim 15m \times 10m \times 4m$, and used 4 different RT60: 0.25 s, 0.5 s, 0.75 s, and 1 s. Consequently, four reverberation evaluation datasets were generated, as presented in Table 1.

4. Experiments

This section examines whether SE front-end co-optimization with our modified training protocols consistently enhances robustness of ResNet18- and LCNN-based CM systems, extending our prior findings [23]. In [23], we trained separate models for each noise type and SNR using pre-generated noise-added data. In contrast, here our experiment trains only two models per network architecture—one for noisy conditions and one for reverberant conditions—using on-the-fly data augmentation to enhance generalization performance. Next, we investigated whether employing pre-trained models alone enhances system robustness. Specifically, we evaluated the performance of spoofing detection systems utilizing W2V2-AASIST and ASR-Conformer, corresponding to the two pre-training approaches shown in Figure 2(c), against noisy and reverberating conditions. Their results were compared to those of AASIST and Conformer CM systems without pre-training. Finally, we implemented SECM-Joint, utilizing the ASR-Conformer pre-trained model as the back-end of the CM system and jointly optimizing it with the proposed DUMENet network for the DSD task. This setup was also designed to evaluate the system's robustness against noise and reverberation. Additionally, we visualized and analyzed the experimental results and conducted ablation studies to further assess the impact of each component.

4.1 Model Parameters and Training Conditions

In the first part of the experiment, the LCNN model, ResNet18 model, and the U-Net-based SE model adopt the same configurations as in our previous work [23]. The embedding size of the CM back-end output is set to 256.

In the second part of the experiments, as shown in Figure 2(c), we evaluated CM systems with pre-training applied to the front-end feature extractor and transfer learning applied to the back-end classifier. For the pre-trained front-end feature extractor approach in Figure 2(c.1), we adopted AASIST, an end-to-end solution proposed by Jung et al. [15], as the baseline model. This model employs a novel heterogeneous stacking graph attention layer that captures artifacts spanning both temporal and spectral domains using a heterogeneous attention mechanism and a stack node. In contrast, the pre-training method introduced by Tek et al. [13] uses the pre-trained wav2vec 2.0 XLS-R model as the front-end, replacing the sinc-layer front-end in AASIST. We refer to this approach as W2V2-AASIST, which has demonstrated superior performance across multiple DSD datasets [1]. For the back-end classifier transfer learning approach shown in Figure 2(c.2), we employed the encoder model of STT En Conformer-CTC Small LibriSpeech model (version 1.0.0) [††] as the pre-trained CM back-end, The encoder model follows the same architecture as that proposed by Gulati et al. [27]. According to the open-source implementation, the convolutional layer in the Conformer-CTC-small-ls model applies a downsampling rate of 1/4. The encoder consists of 13 million parameters, incorporating 4 attention heads and 16 Conformer blocks. The convolutional feature dimension of the encoder is set to 176.

Table 2 shows the parameters of the U-net model used in this work which is the same as our previous work [23]. The total number of blocks is set to 8, with 4 blocks in

[†] https://github.com/LCAV/pyroomacoustics

[††] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_small_ls

Table 2 The parameters of the U-Net based speech enhancement network. **C**(k, s, c) denotes the 2D convolution layer while **TC**(k, s, c) denotes the 2D transposed convolution layer. SE denotes the Squeeze-and-Excitation block here [39]. **EB**x denotes corresponding encoder block x.

Layer name	**Layer Structure**
Conv. 1	**C**(7,1,16)
Encoder Block 1	$\begin{bmatrix} \mathbf{C}(3,1,16) \\ \mathbf{C}(3,1,32) \\ \mathbf{SE} \end{bmatrix} \times 3$
Encoder Block 2	$\begin{bmatrix} \mathbf{C}(3,2,32) \\ \mathbf{C}(3,1,32) \\ \mathbf{SE} \end{bmatrix} \times 4$
Encoder Block 3	$\begin{bmatrix} \mathbf{C}(3,2,64) \\ \mathbf{C}(3,1,64) \\ \mathbf{SE} \end{bmatrix} \times 6$
Encoder Block 4	$\begin{bmatrix} \mathbf{C}(3,1,128) \\ \mathbf{C}(3,1,128) \\ \mathbf{SE} \end{bmatrix} \times 3$
Decoder Block 1	$[\mathit{Concatenate}\ \mathbf{EB}4, C(3,1,32)]$
Decoder Block 2	$[\mathit{Concatenate}\ \mathbf{EB}3, TC(2,1,64)]$
Decoder Block 3	$[\mathit{Concatenate}\ \mathbf{EB}2, TC(2,1,128)]$
Decoder Block 4	$[\mathit{Concatenate}\ \mathbf{EB}1, C(1,1,256)]$
Conv. 2	TC(2 × 1,2 × 1,1)

Table 3 Comparison of the EER% of the ResNet18 and LCNN models evaluated in noise and reverberation conditions in this experiment with the results of previous work [23]. **AVG** denotes the mean EER for different SNRs and RT60s in a given noise or reverberation evaluation data.

Eval. dataset	**SNR**	**Rsenet18**		**LCNN**		**Rsenet18[23]**		**LCNN[23]**	
		Online general noise augmentation				Offline noise data			
		no-SE	U-net	no-SE	U-net	no-SE	U-net	no-SE	U-net
Babble	20 dB	7.08	5.57	7.12	5.00	6.81	3.89	5.69	3.91
	15 dB	9.03	7.80	10.40	7.89	8.63	3.16	8.68	6.10
	10 dB	11.83	10.70	13.57	11.82	9.46	7.39	12.33	8.56
	5 dB	16.68	15.13	18.21	16.76	9.41	8.02	13.00	6.28
	0 dB	24.28	21.46	24.80	23.27	17.47	10.69	17.73	9.30
	AVG	**13.78**	**12.13**	**14.82**	**12.95**	**10.36**	**6.63**	**11.49**	**6.83**
Music	20 dB	6.84	4.65	5.94	3.68	6.95	4.00	6.45	4.13
	15 dB	8.52	5.22	7.80	4.53	8.02	5.37	11.00	6.52
	10 dB	10.58	6.57	10.10	6.16	8.42	7.45	10.59	7.09
	5 dB	14.19	9.04	13.90	9.16	8.85	5.93	11.21	8.59
	0 dB	22.12	14.03	20.69	14.37	15.07	9.99	15.62	9.52
	AVG	**12.45**	**7.90**	**11.69**	**7.58**	**9.46**	**6.55**	**10.97**	**7.17**
Env.	20 dB	6.7	4.98	6.20	3.86	4.98	3.95	6.56	4.48
	15 dB	7.71	5.92	7.11	4.69	6.25	3.59	7.59	5.02
	10 dB	9.72	7.00	8.74	6.30	7.20	6.48	10.02	6.79
	5 dB	13.19	9.01	11.39	9.14	10.79	6.02	13.00	8.24
	0 dB	18.07	13.05	15.53	14.11	16.52	10.41	15.08	10.63
	AVG	**11.08**	**7.99**	**9.79**	**7.62**	**9.15**	**6.09**	**10.45**	**7.03**
	RT60	Online reverberation augmentation				-	-	-	-
		no-SE	U-net	no-SE	U-net	-	-	-	-
Rvb.	0.25 s	6.67	6.47	7.69	6.89	-	-	-	-
	0.5 s	8.71	7.87	8.73	8.43	-	-	-	-
	0.75 s	10.51	9.11	9.84	9.56	-	-	-	-
	1.0 s	12.11	10.02	10.17	10.21	-	-	-	-
	AVG	**9.50**	**8.37**	**9.11**	**8.77**	-	-	-	-

the encoder and 4 blocks in the decoder. The number of channels for each layer in the encoder is set to 16, 32, 64, and 128, respectively. Moreover, the U-Net has one convolution layer and one transposed convolution layer, as mentioned in Chapter 2.2, the proposed DUMENet model use double input features and output the feature mask insted of reconstruct the FBANK features of clean speech.

For feature extraction, except for the AASIST-based CM system, which directly processes raw waveform inputs, all other networks utilise FBANK features as inputs. These features are extracted by applying 80 Mel filters to the spectrogram, which is computed using Hamming windows of 64 ms with a hop size of 8 ms. The duration of all input audio is constrained to 4 seconds, shorter inputs are repeated, while longer inputs are truncated. For training the CM system, the learning rate is initially set to 1e-3. We employ a ReduceLROnPlateau learning rate (LR) scheduler, starting with an initial LR of 0.1. All models are optimised using the Adam optimiser.

In this study, the EER is used as the evaluation metric, which is defined as the threshold at which the false alarm rate (P_{fa}) and the miss rate (P_{miss}) are equal. A lower EER indicates better model performance. To mitigate the variability introduced by random initialisation, we conduct experiments using three different random seeds. Except for the ablation study, the reported experimental results correspond to the lowest EER obtained across the three runs.

4.2 Experimental Results

4.2.1 Comparison of the SE Joint Optimization

We first examine whether our revised training strategy preserves the previously observed conclusion that **joint optimization with an SE front-end improves the robustness of ResNet18- and LCNN-based CM systems**. Table 3 presents the EER results for different models under various noise and reverberation conditions, comparing our findings with prior work [23]. Under all three noise conditions, model performance degrades as the SNR decreases. For the babble noise, ResNet18 with U-net trained with specific matched offline noise type and SNR setting achieves the lowest average EER 6.63%, a 36% (10.36%→6.33%) reduction compared to ResNet18 without U-net. In the on-the-fly noise augmentation setting, ResNet18 jointly optimized with U-net attains an average EER of 12.13%, which, while not as competitive as the matched-noise offline-trained counterpart, still shows an 11.97% (13.78%→12.13%) relative reduction compared to ResNet18 without U-net. This confirms that even under a more generalizable on-the-fly noise augmentation strategy, joint optimization with U-net consistently improves performance, albeit with a slight degradation compared to the matched-noise training. A similar trend is observed in the music noise conditions. When trained on offline matched-noise data, ResNet18 with U-net reduces the average EER by 30.76% (9.46%→6.55%) compared to ResNet18 without U-net. Interestingly, in the on-the-fly general noise augmentation setting, the relative reduction reaches 36.55% (12.45%→7.9%), surpassing that of offline noise training. This suggests that on-the-fly SE allows U-net to generalize better for denoising music noise, further validating the effectiveness of joint optimization in diverse noise scenarios. This conclusion remains consistent for environmental noise, reinforcing the robustness of the proposed approach.

Additionally, we introduce the reverberation augmentation experiments, which reveal a consistent increase in EER with longer reverberation times. Among all models, ResNet18 with U-net achieves the best robustness 8.37% EER. We believe that LCNN's local bias is well-suited to the short-term features of additive noise, whereas ResNet's global modeling capacity better addresses the long-term effects of reverberation and co-optimization with Unet further enhances ResNet's ability to handle longer RT60 conditions. These findings highlight the critical role of U-net in mitigating the impact of both noise and reverberation, enhancing model robustness in complex acoustic environments.

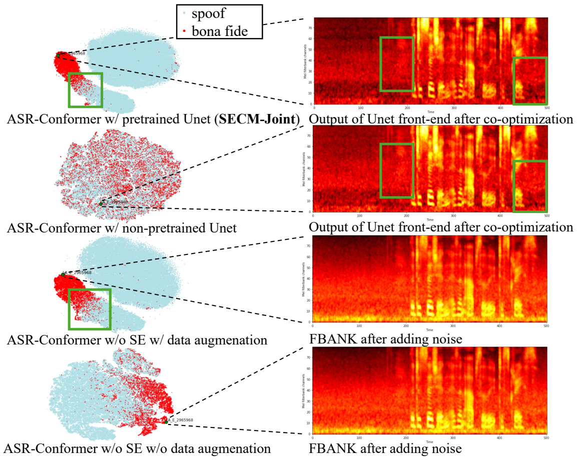
4.2.2 Impact of Pretraining and Transfer Learning on CM System Performance

In the second part of our experiments, we examined how two pre-trained models enhance the robustness of CM systems under various noise and reverberation conditions without data augmentation and enhancement front-end, as shown in Table 4. By comparing models trained on the clean 19LA training set, we aimed to assess the impact of pre-training and transfer learning on robustness. Across all three noise conditions, the introduction of pre-trained models led to a moderate improvement in robustness. However, the improvement was least pronounced under babble noise. This is because babble noise consists of overlapping real speech, which severely interferes with the spoofing detection task at low SNR. Consequently, the performance gains were smaller than those observed under environmental and music noise conditions. Taking W2V2-AASIST as an example, which employs Wav2Vec2 as a front-end pre-trained model, the average EER decreased by 28.79% (24.8%→17.66%) compared to the AASIST baseline, whereas the reductions for music and environmental noise were 68.47% (28.16%→8.88%) and 78.19% (24.07%→5.25%), respectively. Similarly, improvements were also observed in the Conformer-based CM system utilizing ASR transfer learning, although the gains were not as substantial as those from W2V2-AASIST. In terms of absolute EER values, the robustness of the ASR transfer learning-based CM system and the Wav2Vec2 pre-trained CM system is comparable under babble noise. However, under the other two noise conditions, the Wav2Vec2 pre-trained CM system demonstrates superior performance. This can be attributed to the substantial difference in pre-training data, with Wav2Vec2 being trained on approximately 436K hours of data, whereas the ASR pre-trained Conformer model was trained on only 1,000 hours. The disparity in data volume contributes to the performance differences observed in the fine-tuned CM systems.

Under reverberation conditions, the improvements resulting from pre-training varied. The Wav2Vec2-based model exhibited an average EER reduction of 51.27% (52.16%→25.42%), whereas the Conformer-based model showed a reduction of -100.43% (12.84%→25.73%), indicating a significant performance drop. However, at RT60 of 0.25s, Conformer's model also has 49.08% (8.15%→4.15%) lower EER after transfer learning. This indicates that the prior knowledge obtained through ASR transfer learning does not encompass strong reverberation characteristics, leading to severe overfitting in the Conformer model under strong reverberation. This might be because the training data of the ASR conformer does not include high reverberation data. The experimental results in this section indicate that transfer learning with pre-trained models yields modest robustness improvements, but the comparison of Tables 4 and 5 reveals that training data augmentation produces significantly greater gains, demonstrating its critical role for constructing a robust CM system.

Table 4 Comparison of the EER% of CM systems with pre-training components trained on Clean Training data without noise and rererberation augmentation. **AVG** denotes the average EER for different SNRs and RT60s in a given noise or reverberation evaluation data.

Eval. dataset	SNR	Conformer	ASR-Conformer	AASIST	W2V2-AASIST
Babble	20 dB	12.09	8.13	7.97	5.67
	15 dB	14.45	12.39	12.21	10.28
	10 dB	18.05	17.14	23.04	17.23
	5 dB	23.37	22.96	36.04	25.33
	0 dB	28.79	27.38	44.74	29.8
	AVG	**19.35**	**17.60**	**24.80**	**17.66**
Music	20 dB	15.73	11.68	7.44	1.03
	15 dB	20.05	17.85	13.35	2.71
	10 dB	26.99	25.00	26.16	6.05
	5 dB	32.83	31.22	43.44	12.37
	0 dB	37.98	36.34	50.4	22.26
	AVG	**26.72**	**24.42**	**28.16**	**8.88**
Environmental	20 dB	13.09	10.09	9.84	1.28
	15 dB	16.52	14.85	13.27	2.3
	10 dB	20.63	18.93	22.13	3.59
	5 dB	24.62	24.31	33.31	6.98
	0 dB	29.06	28.62	41.78	12.12
	AVG	**20.78**	**19.36**	**24.07**	**5.25**
	RT60				
Reverberation	0.25 s	8.15	4.15	7.37	3.76
	0.5 s	11.04	16.22	50.69	20.52
	0.75 s	14.19	35.01	73.83	34.72
	1.0 s	17.97	47.53	76.76	42.68
	AVG	**12.84**	**26.35**	**52.16**	**25.42**



ASR-Conformer w/ pretrained Unet (**SECM-Joint**) Output of Unet front-end after co-optimization

ASR-Conformer w/ non-pretrained Unet Output of Unet front-end after co-optimization

ASR-Conformer w/o SE w/ data augmenation FBANK after adding noise

ASR-Conformer w/o SE w/o data augmenation FBANK after adding noise

Fig. 4 Visualization of Conformer-based embeddings and features under 10dB environmental noise conditions. (Left) T-SNE plot of output embeddings from Conformer-based detector. (Right) FBANK of sample LA_E_2965968 as input to the Conformer backend.

4.2.3 The Proposed Joint Optimization Method

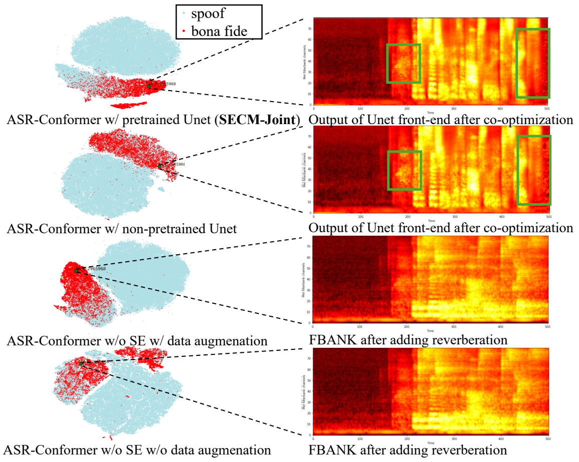
In this part, we integrated ASR-Conformer transfer learning with our proposed DUMENet-based SE model for joint optimisation. However, as shown in Table 5, directly placing the SE module before the pre-trained Conformer led to a decline in CM system robustness across various noise and reverberation conditions. This is primarily because the Unet-based SE front-end was randomly initialised at the beginning of training, making it unable to effectively reconstruct the input FBANK features in the early stages. Since the

Table 5 EER% of Comformer-based CM system before and after using SECM-Joint strategy. **AVG** denotes the average EER for different SNRs and RT60s in a given noise or reverberation evaluation data. **Bold** number indicates the best result for a specific noise or reverberation condition.

Eval. dataset	SNR	Conformer			ASR-Conformer			
		w/o Aug w/o Unet	Noise Aug w/o Unet	Noise Aug w/ Unet	w/o Aug w/o Unet	Noise Aug w/o Unet	Noise Aug w/ Unet	Noise Aug w/ Pre-Unet (**SECM-Joint**)
Babble	20 dB	12.09	7.30	7.37	8.13	5.66	5.93	**4.52**
	15 dB	14.45	8.74	9.20	12.39	8.37	9.61	**7.07**
	10 dB	18.05	**10.70**	11.12	17.14	11.65	14.41	10.97
	5 dB	23.37	**13.99**	14.41	22.96	16.42	22.22	17.23
	0 dB	28.79	**18.86**	18.87	27.38	22.46	31.12	24.09
	AVG	**19.35**	**11.92**	**12.19**	**17.60**	**12.91**	**16.66**	**12.78**
Music	20 dB	15.73	6.76	5.47	11.68	2.62	3.34	**2.43**
	15 dB	20.05	7.70	6.31	17.85	3.06	3.74	**2.82**
	10 dB	26.99	9.08	7.48	25.00	3.87	4.64	**3.67**
	5 dB	32.83	11.68	10.16	31.22	5.71	6.65	**5.45**
	0 dB	37.98	16.75	14.17	36.34	**10.17**	11.25	10.4
	AVG	**26.72**	**10.39**	**8.72**	**24.42**	**5.09**	**5.92**	**4.95**
Environmental	20 dB	13.09	6.61	5.67	10.09	2.84	3.28	**2.57**
	15 dB	16.52	7.52	6.27	14.85	3.13	3.66	**2.86**
	10 dB	20.63	8.13	7.18	18.93	3.72	4.24	**3.46**
	5 dB	24.62	9.71	8.65	24.31	4.68	5.14	**4.32**
	0 dB	29.06	12.96	11.15	28.62	6.85	7.45	**6.79**
	AVG	**20.78**	**8.91**	**7.78**	**19.36**	**4.24**	**4.75**	**4.00**
	RT60	w/o Aug w/o Unet	Rvb. Aug w/o Unet	Rvb. Aug w/ Unet	w/o Aug w/o Unet	Rvb. Aug w/o Unet	Rvb. Aug w/ Unet	Rvb. Aug w/ Pre-Unet (**SECM-Joint**)
Reverberation	0.25 s	8.15	7.76	8.22	4.15	5.91	5.97	**5.38**
	0.5 s	11.04	9.04	9.44	16.22	7.18	7.37	**6.94**
	0.75 s	14.19	10.39	10.12	35.01	8.14	8.46	**7.87**
	1.0 s	17.97	11.19	10.84	47.53	8.78	9.35	**8.43**
	AVG	**12.84**	**9.60**	**9.66**	**26.35**	**7.50**	**7.79**	**7.16**

Table 6 EER% of ablation experiments on the proposed DUMENet.

Test Model		Babble (SNR)					Music (SNR)					Environmental (SNR)					Reverberation (RT60)			
		20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB	0.25 s	0.5 s	0.75 s	1.0 s
SECM-Joint	**DUMENet**	4.52	**7.07**	**10.97**	**17.23**	**24.09**	**2.43**	**2.82**	**3.67**	**5.45**	10.4	**2.57**	**2.86**	**3.46**	**4.32**	**6.79**	**5.38**	**6.94**	**7.87**	**8.43**
	->w/o mask	**4.49**	7.12	11.31	17.73	24.94	2.61	3.03	3.89	5.81	**10.36**	2.80	3.01	3.64	4.53	6.85	6.1	7.61	8.55	9.21
	->w/o Multi input	5.63	9.18	14.33	22.18	30.07	3.03	3.21	4.23	6.15	10.56	2.99	3.24	3.85	4.79	6.78	5.88	7.98	9.00	9.65



ASR-Conformer w/ pretrained Unet (**SECM-Joint**) Output of Unet front-end after co-optimization

ASR-Conformer w/ non-pretrained Unet Output of Unet front-end after co-optimization

ASR-Conformer w/o SE w/ data augmentation FBANK after adding reverberation

ASR-Conformer w/o SE w/o data augmentation FBANK after adding reverberation

Fig. 5 Visualization of Conformer-based embeddings and features under RT60 1s reverberation conditions. (Left) T-SNE plot of output embeddings from Conformer-based detector. (Right) FBANK of sample LA_E_2965968 as input to the Conformer backend.

Conformer model had already been pre-trained, its expected input was structured FBANK features, and the mismatch between the Unet's output and the Conformer's expected input domain increased the difficulty of training. To address this issue, we pre-trained the Unet front-end for 200 epochs and continued to optimize its parameters jointly with the ASR-Conformer model. The results after pre-training, shown in the last column of Table 5, indicate that compared to data augmentation without transfer learning, the average EER reductions for music, environmental noise, and reverberation were 52.38% (10.39%→4.95%), 55.12% (8.91%→4.00%), and 25.38% (9.60%→7.16%), respectively. This demonstrates that the SECM-Joint approach effectively combines the benefits of pre-trained Conformer models and SE models, leading to improved robustness against noise and reverberation compared to either method alone. However, under babble noise conditions, the average EER increased by 7.23% (11.92%→12.78%), indicating that general-purpose SE models struggle with babble noise. This is particularly evident at extremely low SNRs, where the SE model fails to distinguish between target speech components and background speech-like noise, ultimately degrading CM system performance. Based on these findings, we hypothesise that using a speaker-specific target speaker extraction front-end, incorporating with speaker embeddings, could mitigate this issue, which we plan to explore in future research.

Figure 4 and Figure 5 present the T-SNE visualizations of the embeddings produced by the ASR transfer learning-based Conformer, along with the Unet-reconstructed FBANK outputs of the LA_E_2965968 sample under four different training conditions, namely SECM-Joint, ASR-Conformer with non-pretrained Unet, ASR-

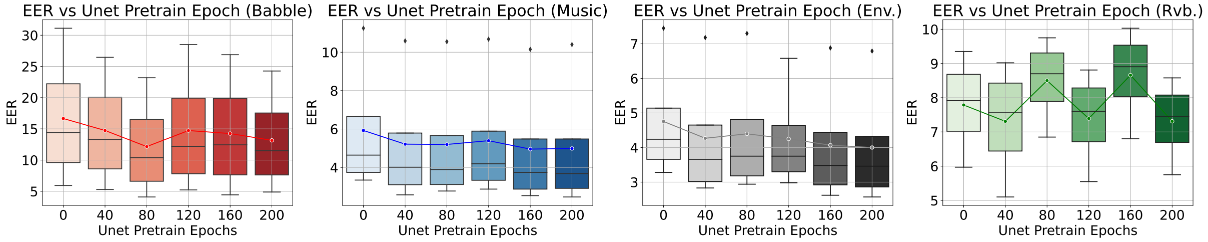


Fig. 6 Comparison of the performance of Unet-based SE model pre-training epochs on SECM-Joint CM performance. Line traces average EER values of each box plot.

Conformer without SE with data augmenation, and ASR-Conformer without SE without data augmenation.

From the visual analysis, it can be observed that without pre-training, the Unet model exhibits less distinguishable embedding distributions in the T-SNE visualization, making classification more challenging. Compared to the embeddings obtained through systems without speech enhancement frontend but trained with data augmentation alone, the joint optimization with DUMENet slightly enhances separability. Additionally, by comparing the Unet-reconstructed FBANK output with the noise-added LA_E_2965968 test sample, we observe that the noise reduction effect achieved through joint optimization closely resembles that of the pre-trained Unet. However, when DUMENet is not pre-trained, the reconstructed output contains more residual noise artifacts, which primarily localized at speech onset/offset regions, as highlighted by green bounding boxes in Figure 4 and 5.

4.2.4 The DUMENet

In the previous section, we noted that if the Unet model is not pre-trained, its output becomes misaligned with the expected input of the ASR-Conformer, leading to the anomalous performance degradation observed earlier. To further investigate this, we conducted experiments to examine the relationship between the number of pre-training epochs of the Unet-based SE model and the final performance of the SECM-Joint CM system. The results, presented in Figure 6, indicate that for the three types of noise conditions, the EER decreases significantly as the number of Unet pre-training epochs increases. However, under reverberation conditions, the EER values for 40 and 200 pre-training epochs are nearly identical. This suggests that since the ASR-Conformer model has not been trained with reverberant data, it may not be very sensitive to the level of reverberation in the input FBANK features. Furthermore, we conducted an ablation study on the proposed DUMENet model by comparing its performance with two alternative configurations: one without the mask mechanism and another that additionally removes the dual-branch input. The results, summarized in Table 6, demonstrate that under most conditions, DUMENet outperforms both modified Unet configurations.

5. Conclusion

In this study, we address the challenge of enhancing the robustness of speech anti-spoofing countermeasure (CM) systems under noisy and reverberant conditions by introducing a transfer learning-based Speech Enhancement Counter Measure Joint optimization approach, SECM-Joint. Experimental results demonstrate that, except for the high-intensity babble noise condition, SECM-Joint approach significantly reduces the Equal Error Rate (EER) across various acoustic conditions, including music and Environmental noise with different Signal-to-Noise Ratios (SNRs). Compared to the Conformer baseline model without pre-training and relying solely on data augmentation, SECM-Joint achieves an EER reduction ranging from 19.11% to 64.05% under noisy conditions and from 23.23% to 30.67% under different RT60 reverberation scenarios.

These findings suggest that SECM-Joint and DUMENet improve the adaptability of CM systems in complex acoustic environments, making them promising for real-world applications where robustness to environmental variability is critical. However, our study was limited to specific noise types and reverberation levels; future work could explore a broader range of acoustic conditions and further refine the SECM-Joint framework. Additionally, we observed that under babble noise conditions, both bona fide and spoofed samples retain real speech characteristics, posing a significant challenge for spoofing detection. Future research could investigate this issue in greater depth, for instance, by incorporating the SE module with a speaker-specific extraction module to better address this challenge.

References

[1] J. Yi, C. Wang, J. Tao, X. Zhang, C. Zhang, and Y. Zhao, "Audio Deepfake Detection: A survey," in *arXiv preprint arXiv:2308.14970*, 2023.

[2] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z. H. Tan, "Further Optimisations of Constant Q Cepstral Processing for Integrated Utterance and Text-dependent Speaker Verification," in *Proc. of SLT*, pp. 179–185, 2016.

[3] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. of Interspeech*, pp. 1033–1037, 2019.

[4] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Proc. of Interspeech*, pp. 1078–1082, 2019.

[5] F. Chen, S. Deng, T. Zheng, Y. He, and J. Han, "Graph-Based Spectro-Temporal Dependency Modeling for Anti-Spoofing," in *Proc. of ICASSP*, pp. 1–5, 2023.

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Trans. on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[7] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J.

Yamagishi, and K. A. Lee, “The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection,” in *Proc. of Interspeech*, pp. 2–6, 2017.

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection,” *Proc. of Interspeech*, pp. 1008–1012, 2019.

[9] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, “Learning from Yourself: A Self-Distillation Method for Fake Speech Detection,” *Proc. of ICASSP*, pp. 1–5, 2023.

[10] X. Wang and J. Yamagishi, “A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection,” *Proc. of Interspeech*, pp. 4259–4263, 2021.

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A Comparison of Features for Synthetic Speech Detection,” *Proc. of Interspeech*, pp. 2087–2091, 2015.

[12] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, “Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion,” *Proc. of Interspeech*, pp. 77–81, 2018.

[13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end Anti-spoofing with Rawnet2,” in *Proc. of ICASSP*, pp. 6369–6373, 2021.

[14] H. Tak, J. Weon Jung, J. Patino, M. R. Kamble, M. Todisco, and N. W. D. Evans, “End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection,” *Proc. of ASVspoof Workshop*, pp. 1–8, 2021.

[15] J. Weon Jung, H.-S. Heo, H. Tak, H. Jin Shim, J. S. Chung, B.-J. Lee, H. Jin Yu, and N. W. D. Evans, “AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks,” *Proc. of ICASSP*, pp. 6367-–6371, 2022.

[16] X. Liu, M. Liu, L. Wang, K.-A. Lee, H. Zhang, and J. Dang, “Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection,” *Proc. of ICASSP*, pp. 1–5, 2023.

[17] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” *Proc. of ASVspoof Workshop*, pp. 47–54, 2021.

[18] H. Tak, M. R. Kamble, J. Patino, M. Todisco, and N. W. D. Evans, “Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing,” *Proc. of ICASSP*, pp. 6382–6386, 2022.

[19] Y. Wang, X. Wang, H. Nishizaki, and M. Li, “Low Pass Filtering and Bandwidth Extension for Robust Anti-Spoofing Countermeasure Against Codec Variabilities,” *Proc. of ISCSLP*, pp. 438–442, 2022.

[20] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, “Spoofing Detection Goes Noisy: An Analysis of Synthetic Speech Detection in the Presence of Additive Noise,” *Speech Communication*, vol. 85, pp. 83–97, 2016.

[21] H. Yu, A. Sarkar, D. A. L. Thomsen, Z.-H. Tan, Z. Ma, and J. Guo, “Effect of Multi-condition Training and Speech Enhancement Methods on Spoofing Detection,” *Proc. of SPLINE*, pp. 1–5, 2016.

[22] C. Fan, M. Ding, J. Tao, R. Fu, J. Yi, Z. Wen, and Z. Lv, “Dual-Branch Knowledge Distillation for Noise-Robust Synthetic Speech Detection,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 2453-2466, 2024.

[23] X. Wang, B. Zeng, S. Hongbin, Y. Wan, and M. Li, “Robust Audio Anti-spoofing Countermeasure with Joint Training of Front-end and Back-end Models,” *Proc. of Interspeech*, pp. 4004–4008, 2023.

[24] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, “Fifty Years of Artificial Reverberation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[25] J.-H. Kim, J. Heo, H.-J. Shim, and H.-J. Yu, “Extended U-Net for Speaker Verification in Noisy Environments,” *Proc. of Interspeech*, pp. 590–594, 2022.

[26] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*, 2nd ed., ch. 4, sec. 4, 2000.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” *Proc. of Interspeech*, pp. 5036–5040, 2020.

[28] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook et al., “Nemo: a Toolkit for Building AI Applications Using Neural Modules,” arXiv:1909.09577, 2019.

[29] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. Lee, and H. Meng, “MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification,” *Proc. of Interspeech*, pp. 306–310, 2022.

[30] D. Cai and M. Li, “Leveraging ASR Pretrained Conformers for Speaker Verification Through Transfer Learning and Knowledge Distillation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 3532–3545, 2024.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. of the CVPR*, pp. 770–778, 2016.

[32] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio, “Maxout Networks,” *Proc. of ICML*, pp. 1319–1327, 2013.

[33] M. Schuster and K. K. Paliwal, “Bidirectional Recurrent Neural Networks,” *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[34] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, Hanilçi, M. Sahidullah, and A. Sizov, “ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge,” *Proc. of Interspeech*, pp. 2037–2041, 2015.

[35] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr Vctk Corpus: English Multi-speaker Corpus for Cstr Voice Cloning Toolkit (version 0.92),” 2019.

[36] D. Snyder, G. Chen, and D. Povey, “Musan: A Music, Speech, and Noise Corpus,” *arXiv preprint arXiv:1510.08484*, 2015.

[37] Y. Luo and R. Gu, “Fast Random Approximation of Multi-channel Room Impulse Response,” *Proc. of ICASSP Workshop*, pp. 449–454, 2024.

[38] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition,” *Proc. of ICASSP*, pp. 5220–5224, 2017.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation Networks,” *Proc. of CVPR*, pp. 7132–7141, 2018.

[40] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, “ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” *Proc. of ASVspoof Workshop*, vol. 13, 2019.

Yikang Wang received his B.Eng. degree in Electrical Engineering and Automation from Shandong Jianzhu University in 2018, and his M.Eng. degree in Mechatronics from the University of Yamanashi in 2021. He is currently pursuing a Ph.D. degree in System Integration Engineering at the University of Yamanashi. His research interests include sound classification, acoustic feature improvement, and deepfake speech detection.

Xingming Wang received his B.Eng. degree in Computer Science from Wuhan University (China) in 2020, and his M.Eng. degree in Computer Science from the same university in 2023. His research focused on audio anti-spoofing, speaker verification, and language identification.

Chee Siang Leow received his B.E., M.E. and Ph.D. degrees in mechatronics engineering from University of Yamanashi in 2018, 2020, and 2024 respectively. He is currently assistant professor at Graduate Faculty of Interdisciplinary Research Faculty of Engineering, University of Yamahashi, since 2024 April. His main research interests include the application of artificial intelligence technology in speech recognition and computer vision. He is a member of Information Processing Society of Japan (IPSJ).

Qishan Zhang Qishan Zhang is currently pursuing the degree with the College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi, China. He research interests include audio deepfake detection and anti-spoofing.

Ming Li Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in 2013. He is currently a Professor of Electronical and Computer Engineering at Duke Kunshan University. He is also an Adjunct Professor at School of Computer Science in Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavior signal processing. He is also a senior member of IEEE.

Hiromitsu Nishizaki received his B.E., M.E., and Ph.D. degrees in engineering from Toyohashi University of Technology, Japan in 1998, 2020, and 2003, respectively. From August 2015 to March 2016, he was a visiting researcher at the National Taiwan University in the Republic of China. He has been a professor at Graduate School of Interdisciplinary Research, University of Yamahashi, since 2022. His research interests include spoken language processing and image processing using deep learning. He is also a senior member of the Institute of Electronics, Information, and Communication Engineers (IEICE) and IEEE.