<Society logo(s) and publication title will appear here.>

Multi-channel Replay Speech Detection using an Adaptive Learnable Beamformer

Michael Neri[1], Member, IEEE, Tuomas Virtanen[1], Fellow, IEEE

[1]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, 33720 Finland

Corresponding author: Michael Neri (email: michael.neri@tuni.fi).

ABSTRACT Replay attacks belong to the class of severe threats against voice-controlled systems, exploiting the easy accessibility of speech signals by recorded and replayed speech to grant unauthorized access to sensitive data. In this work, we propose a multi-channel neural network architecture called M-ALRAD for the detection of replay attacks based on spatial audio features. This approach integrates a learnable adaptive beamformer with a convolutional recurrent neural network, allowing for joint optimization of spatial filtering and classification. Experiments have been carried out on the ReMASC dataset, which is a state-of-the-art multi-channel replay speech detection dataset encompassing four microphones with diverse array configurations and four environments. Results on the ReMASC dataset show the superiority of the approach compared to the state-of-the-art and yield substantial improvements for challenging acoustic environments. In addition, we demonstrate that our approach is able to better generalize to unseen environments with respect to prior studies.

INDEX TERMS Replay attack, Physical Access, Beamforming, Spatial Audio, Voice anti-spoofing.

I. Introduction

In recent years, the use of voice assistants (VAs) has been increasing for human-machine interaction, employing voice as a biometric trait [1]. With VAs, it is possible to control smart devices in our homes using Internet of things (IoT) networks and transmit sensitive information over the Internet. For this reason, malicious users started to attack audio-based data to grant unauthorized access to personal data and devices.

Generally, there are several types of attacks that a malicious user can exploit toward an audio recording [2]. It is possible to perform a logical access (LA) attack, i.e., applying text-to-speech (TTS) and/or voice conversion (VC) algorithms to modify the speech content or imitate the user. Lately, the deepfake (DF) attack showed how compression and quantization operations to audio recordings can hide LA artifacts.

Differently, with physical access (PA) attacks the objective is to counterfeit the automatic speaker verification (ASV) system at the microphone level. In this category, there are *impersonation attacks* [1] and *replay attacks* [3]. The former involves an attacker who is able to mimic the user speech of the target to grant access through the ASV system. The latter, instead, consists of a fraudulent recording of the genuine
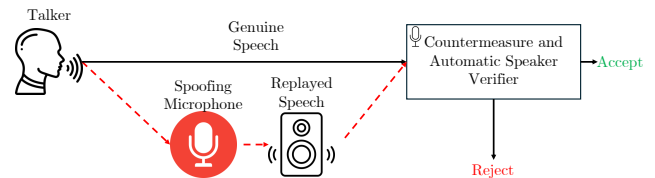
FIGURE 1. **The problem of detecting the replay speech attack. In this scenario, a genuine speech from a talker can be either directly transmitted to a biometric system or intercepted by a spoofing microphone, recorded, and replayed through a speaker. The replayed speech, captured by the system, may be misclassified as genuine, granting unauthorized access to private data. A countermeasure and automatic speaker verification module are typically used to distinguish between genuine speech and spoofed (replayed) speech.**

speaker using a spoofing microphone, which is later replayed to the ASV microphone. An example of a replay attack is shown in Figure 1.

The research community has mainly focused its attention on LA attacks by developing countermeasure (CM) models that are responsible for detecting synthethic generation artifacts. In fact, several datasets have been published with the aim to tackle these attacks, e.g., the ASVSpoof series [2], [4]. However, voice is different to other biometric markers

as it is always publicly exposed, easing the collection of spoofed recordings by attackers from the target speaker [5]. This characteristic allows malicious users to easily perform PA even without any professional audio expertise. In fact, *replay attacks* can be performed by using commercial-off-the-shelf (COTS) devices like smartphone and loudspeakers with high success rates.

In the literature, several works have tried to mitigate this type of attack by providing single-channel speech datasets, such as RedDots [6], ASVSpoof2017 PA [7], ASVSpoof2019 [8] PA, and ASVSpoof2021 PA [2]. Building upon these datasets, researchers have been deploying various methods to address single-channel replay attack detection, such as capsule networks [9], convolutional neural networks (CNNs) with hand-crafted audio features combined with time-frequency representations [10]–[12], and teacher-student architectures to reduce the computational complexity for real-time detection [13].

However, for speech enhancement and separation tasks, microphone arrays are usually employed in ASV systems to exploit spatial information and improve audio quality [14]. Moreover, multi-channel data can be beneficial for detecting the replay detection for several reasons: (i) multi-channel recordings encompass audio spatial cues that can help the detection [3], [15], and (ii) this spatial information cannot be easily counterfeited by an attacker, differently to single-channel data where temporal and frequency cues can be manipulated to fool an ASV system [16]. Unfortunately, single-channel replay detectors do not exploit spatial information from multi-channel recordings. In this regard, single-channel replay detectors demonstrated poor generalization capabilities across different environments and devices on multi-channel data [3]. Moreover, there is a lack of datasets and models that are capable to collect and detect replay attacks based on multi-channel recordings. Realistic Replay Attack Microphone Array Speech (ReMASC) is the only publicly available dataset that encompasses real multi-channel audio samples, including different microphone arrays, playback devices, and environments. The baseline of this dataset was CQCC-GMM [15], which is a Gaussian mixture model (GMM) classifier trained on constant-Q cepstral coefficients (CQCCs). The authors of the database also proposed a convolutional recurrent neural network (CRNN) which acts as both a learnable and adaptive time-domain beamformer with a CRNN replay attack classifier [3].

Recent works in state-of-the-art (SOTA) focused on hand-crafted single-channel features with simple classifiers (such as CNNs and GMMs) after a beamforming phase (TECC [17], CTECC [18], and ETECC [19]). However, all these methods suffer from generalization capabilities, i.e., changing the acoustic properties on the test set yields random guess predictions. This problem also occurs with single-channel utterances [11]. In addition, performance for known environments is still not enough for production stage.

To tackle the aforementioned problems, the contributions of this work are:

- The definition of a novel **M**ulti-channel **R**eplay **A**ttack **D**etector that exploits a **L**earnable time-frequency **A**daptive beamformer to classify between genuine and spoofed audio recordings, namely **M-ALRAD**. It includes a time-frequency beamformer and a CRNN classifier.
- We define a beamformer regularization during the training process to enhance the generalization capabilities of M-ALRAD to unseen environments.
- Comparisons with the state-of-the-art show the effectiveness of the proposed approach in all the microphone configurations. In addition, we demonstrate that our approach is able to better generalize in environments where no a-priori knowledge is available about acoustic properties with respect to SOTA methods.

The paper is organized as follows. Section II introduces the proposed multi-channel replay attacks, illustrating the learnable beamformer, the deep neural network (DNN) that performs the classification, and the dataset under-analysis. Then, experimental results are shown in Section III, comparing different DNN configurations and SOTA approaches. Finally, possible future works and conclusions are drawn in Section IV.

II. Proposed method

In this section, the architecture of M-ALRAD is depicted. The structure of the section is as follows. First, the problem statement of this paper is shown. Then, we introduce how multi-channel recordings are processed by the adaptive learnable beamformer shown in Figure 2. Next, we describe the DNN that is responsible for the replay attack detection. Finally, we detail how the model is optimized during the training process.

A. Problem statement
Given a recorded speech from an array with N microphones $\mathbf{x}[n] = \{x_i[n], 0 \leq i < N\}$, the aim of this work is to identify whether the speech is genuine, i.e., produced by an human speaker, or spoofed, i.e., recorded from the target talker and then replayed again by using a loudspeaker.

B. Learning-based adaptive beamforming
Generally, beamformers are typically designed using objectives such as minimum variance distortionless response (MVDR) [20] or multichannel Wiener filtering (MWF) [21], which are optimized to minimize noise power or mean squared error, while preserving the desired signal. While this approach suits tasks like speech enhancement and separation, where less distortion and noise improve performance, it can undermine replay attack detection by removing key cues in noisy or high-frequency components. As done in [3], a more effective solution is to jointly optimize the beamformer with

<Society logo(s) and publication title will appear here.>
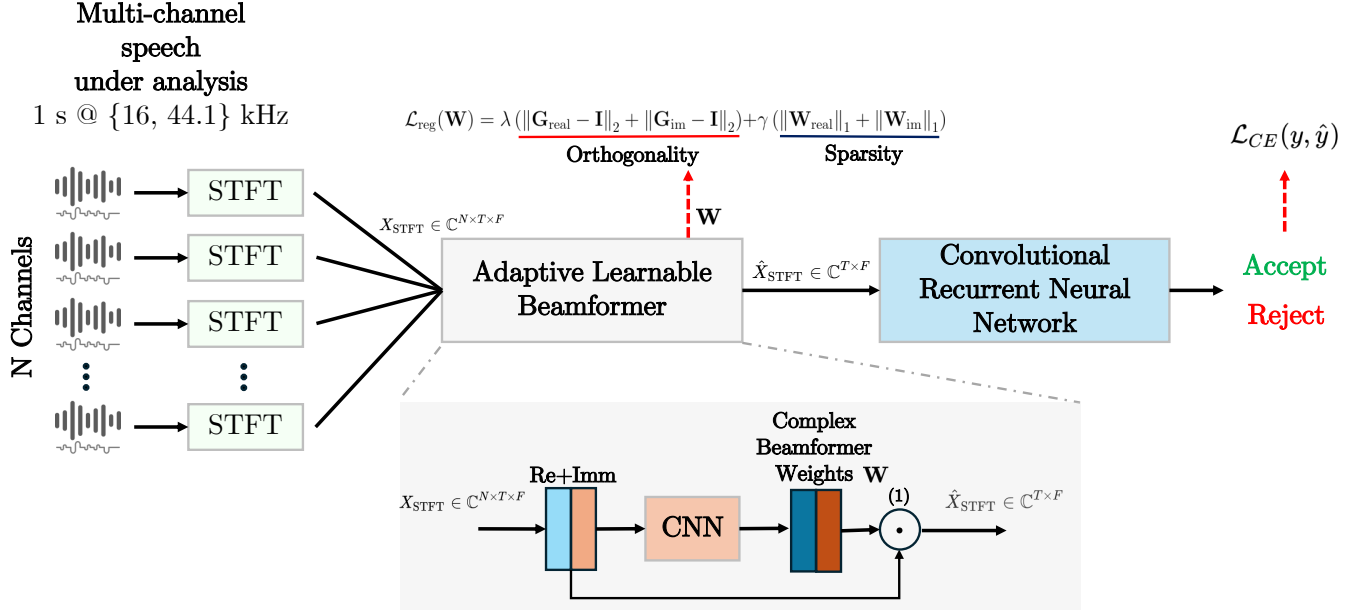
FIGURE 2. Overview of M-ALRAD. First, a multi-channel recording $\mathbf{x}[n]$ is channel-wise analyzed by means of STFT to extract complex spectrograms. Then, the adaptive beamformer extracts a set of complex weights $\mathbf{W}$ that are used to obtain the final beamformed single-channel spectrogram $\hat{X}_{\mathrm{STFT}}$. Finally, a CRNN is employed to decide whether the recording is genuine or spoofed.

the replay attack detector by means of the back-propagation algorithm.

In our method, we design a learning-based adaptive beamformer. Specifically, the module is a CNN that analyzes complex STFTs of the input multi-channel recordings to output a single-channel beamformed complex STFT, i.e., $f_{BM} : \mathbb{C}^{N \times T \times F} \rightarrow \mathbb{C}^{T \times F}$, where T and F denote the number of time and frequency bins, respectively. Next, a CNN is defined to process all the complex spectrograms and to output the beamforming weights $\mathbf{W} \in \mathbb{C}^{N \times T \times F}$. The CNN is a sequence of Conv2D-BatchNorm-ELU-Conv2D operations with n_f 3×3 filters. Real and imaginary parts are concatenated along the channel dimension, both at the input and at the output of the learnable beamformer. Specifically, the CNN outputs real and imaginary parts, namely $\mathbf{W}_{\mathrm{re}} \in \mathbb{R}^{N \times T \times F}$ and $\mathbf{W}_{\mathrm{im}} \in \mathbb{R}^{N \times T \times F}$ respectively, which are combined to compose the complex weights as $\mathbf{W} = \mathbf{W}_{\mathrm{re}} + j\mathbf{W}_{\mathrm{im}}$. Finally, the beamformed single-channel complex spectrogram $\hat{X}_{\mathrm{STFT}} \in \mathbb{C}^{T \times F}$ is obtained by summing all the input STFTs multiplied by the beamforming weights as

$$\hat{X}_{\mathrm{STFT}_{t,f}} = \sum_n X_{\mathrm{STFT}_{n,t,f}} \cdot w_{n,t,f}. \tag{1}$$

C. Convolutional recurrent neural network classifier
Given the single-channel complex spectrogram from the beamformer, we employ a CRNN that has been already used for speaker distance estimation from single-channel recordings [22], [23]. Specifically, magnitude and phase of the beamformed STFT are computed. We extract the sin&cos features for each time-frequency point of the STFT

raw phase, i.e., $\sin(\angle\mathrm{STFT}\{\mathbf{x}\})$ and $\cos(\angle\mathrm{STFT}\{\mathbf{x}\})$. The features are then arranged in a $T \times F \times 3$ tensor to be processed by three convolutional layers. Each convolutional block consists of a 2D convolutional layer containing $P_i = \{P_1, P_2, P_3\} = \{32, 64, 128\}$ 1×3 filters, where i corresponds to the i-th layer, followed by batch normalization. Max and average pooling operation are computed in parallel along the frequency dimension in order to be summed. The applied activation function is the exponential linear unit (ELU) [24]. The max and average pooling rate of each layer is $MP = \{MP_1, MP_2, MP_3\} = \{8, 8, 4\}$. Two bi-directional gated recurrent unit (GRU) layers, with $\tanh(\cdot)$ as activation function, are applied to the feature maps from the convolutional layers. In this implementation, each GRU has 128 neurons. Finally, the hidden state of the last time step from the GRU output is extracted, which represents the final summary of the entire sequence, and converted into the prediction $\hat{y} \in \mathbb{R}$ by means of a fully connected layer.

D. Beamformer regularization and classification loss
To increase the generalization capabilities of the proposed approach, we propose to enforce complex orthogonality and sparse weights in the learnable beamformer.

Specifically, since the matrix W is 3-dimensional, we decompose the complex weights $\mathbf{W} \in \mathbb{C}^{N \times T \times F}$ into real and imaginary parts, reshaping them into $\mathbf{W}_{\mathrm{re}} \in \mathbb{R}^{N \times F \times T}$ into $\mathbb{R}^{N \times FT}$, and similarly for $\mathbf{W}_{\mathrm{im}}$

$$\mathbf{W} = \mathbf{W}_{\mathrm{re}} + j\mathbf{W}_{\mathrm{im}}, \tag{2}$$

and enforce orthogonality independently on both parts:

$$\mathcal{L}_{\text{ortho}} = \left\| \mathbf{W}_{\text{re}} \mathbf{W}_{\text{re}}^T - \mathbf{I} \right\|_2 + \left\| \mathbf{W}_{\text{im}} \mathbf{W}_{\text{im}}^T - \mathbf{I} \right\|_2, \quad (3)$$

where $\mathbf{I}$ is the identity matrix. This is done by computing their Gram matrices

$$\mathbf{G}_{\text{re}} = \mathbf{W}_{\text{re}} \mathbf{W}_{\text{re}}^T, \qquad \mathbf{G}_{\text{im}} = \mathbf{W}_{\text{im}} \mathbf{W}_{\text{im}}^T. \quad (4)$$

Finally, the full regularization loss, including the sparsity constraint, is:

$$\mathcal{L}_{\text{reg}}(\mathbf{W}) = \lambda \left(\|\mathbf{G}_{\text{re}} - \mathbf{I}\|_2 + \|\mathbf{G}_{\text{im}} - \mathbf{I}\|_2 \right) + \\ \gamma \left(\|\mathbf{W}_{\text{re}}\|_1 + \|\mathbf{W}_{\text{im}}\|_1 \right). \quad (5)$$

where $||\cdot||_p$ is the p-norm, and $\{\lambda, \gamma\}$ are used to combine losses with different magnitudes. The last term imposes sparsity of beamformer's weights, encouraging stability and prevent overly large or noisy weights in the learnable beamformer.

Finally, the overall architecture is trained to distinguish between genuine and replay recording by minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}(y, \hat{y})$ between ground truth y and prediction $\hat{y}$ in conjunction with the beamformer regularizer

$$\mathcal{L}(y, \hat{y}, \mathbf{W}) = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \mathcal{L}_{reg}(\mathbf{W}). \quad (6)$$

E. Implementation details
Following [3], we only analyze the first 1 second of multi-channel recordings to reduce the computational complexity of the approach, as increasing the analysis segment length to 1.5 or 2 seconds was not found to improve the performance. Separate models are trained for each microphone array. Sampling rate for microphones $\{D1, D2, D3\}$ is 44.1 kHz, whereas for D4 it is 16 kHz. The STFT is computed with a Hanning window of length 32 and 46 ms with 50% overlap for 44.1 and 16 kHz, respectively. The number of intermediate channels in the beamformer CNN is $n_f = 64$. The proposed model is then trained with a batch size of 32 using a cosine annealing scheduled learning rate of 0.001 for 50 epochs. We set $\lambda = \gamma = 0.00001$. These hyperparameters have been selected based on a validation set randomly extracted from the training set with a 90-10 split ratio. We re-weight the cross-entropy loss for each class using the normalized reciprocal of the sample number of the class in the training set to avoid the class-imbalance problem. M-ALRAD is trained for approximately 50 minutes on a NVIDIA GeForce RTX 4070. The average measured inference time of a single multi-channel recording is around 10 ms. On average, M-ALRAD consists of 300k learnable parameters. A version of the model is trained for each microphone array.

III. Experimental results
In this section, we provide the results of our approach on ReMASC dataset. In more detail, we compare the microphone-wise performance of our approach with the ReMASC baseline. Then, we analyze the effect of varying the acoustical properties of the scenario. Finally, a comparison with prior works and an ablation study are provided.

A. Evaluation
The ReMASC is used to assess the proposed approach [15]. The dataset features a wide range of speakers, with variations in gender and vocal characteristics, providing diversity in the multi-channel audio data. It also incorporates recordings from four environments, including one outdoor scenario (Env-A), two enclosures (Env-B and Env-C) and one moving vehicle (Env-D), to capture different acoustic properties. Two spoofing microphones are employed for recording the genuine speech and then replayed using four different playback devices, including diverse quality. All recordings are synchronized across 4 ASV microphones $\{D1, D2, D3, D4\}$ with 2, 4, 6, and 7 channels, respectively. In total, there are $9,240$ and $45,472$ genuine and replay audio samples, respectively.

To assess the performance of our approach and directly compare with SOTA models, equal error rate (EER) metric is used following the same train-test split provided by the authors of the dataset [15]. To reduce the stochastic behavior of training the proposed DNN, five independent runs are collected to compute the 95% mean confidence intervals.

B. Results and comparison with baseline
Table 1 shows the microphone-wise performance comparison to ReMASC baseline NN-Multichannel [3]. In addition, to evaluate the effect of spatial information, we provide the results for the case where a single-channel (the first of the array) is replicated to each microphone channel and then fed to the proposed DNN, namely ALRAD. The direct comparison between M-ALRAD and NN-Multichannel shows that our approach achieves better results across all the microphones. Notably, an average decrease of 5% in terms of EER can be highlighted, except for D1 where the improvement is approximately 10%.

As noted in [3], the use of multiple channels proves advantageous for the detection of replay attacks. However, in accordance with previous studies [3], [15], [19], it is important to note that recognizing attacks becomes more challenging with arrays that have more microphones. Specifically, while using both channels of a binaural microphone improves performance compared to a single channel, using arrays with more microphones yields worse results. To better analyze this behavior, receiving operating characteristic (ROC) curves of the proposed approach for each microphone are depicted in Figure 3.

C. Generalization across unseen environments
To assess the generalization capabilities of our approach, we carried out dependent training procedures as in [15]. Specifically, we assess the performance of the model when either samples of the target environment are present (*environment-dependent*) or not (*environment-independent*) in the training
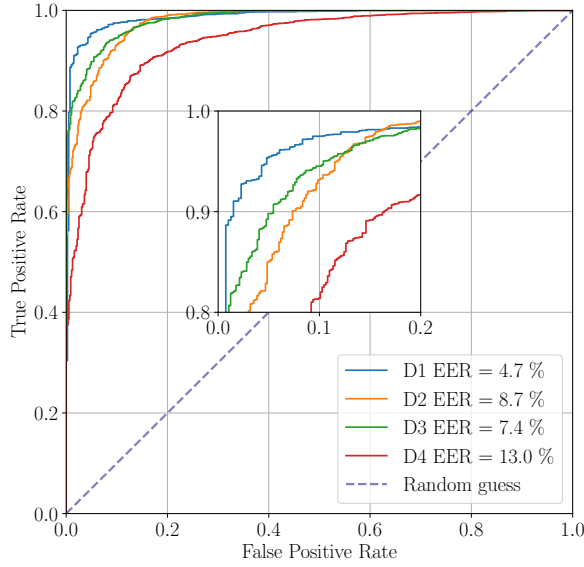
FIGURE 3. ROC curves for each microphone's model.

TABLE 1. Microphone-wise comparison of EER (%) with ReMASC baseline.

Methods	D1	D2	D3	D4
NN-Single [3]	16.6	23.7	23.7	27.5
NN-Dummy Multichannel [3]	16.0	23.2	24.5	25.2
NN-Multichannel [3]	14.9	15.4	16.5	19.8
ALRAD	5.5 ± 2.6	11.9 ± 2.2	19.5 ± 2.2	21.7 ± 2.1
M-ALRAD	$\mathbf{5.2} \pm 1.2$	$\mathbf{10.0} \pm 0.9$	$\mathbf{10.4} \pm 2.1$	$\mathbf{14.2} \pm 0.8$

set. For example, results for *environment-dependent* Env-A are obtained by training M-ALRAD for each microphone with all the environments, testing on Env-A, and averaging the EERs. Regarding *environment-independent* on the same environment, Env-A is not present in the training set, enabling the evaluation of M-ALRAD to unseen scenes.

Table 2 shows a comparison between SOTA approaches and our replay attack detector. Generally, in the case when the model has some a-priori knowledge of the environment, our approach outperforms the SOTA approaches in the outdoor scenario and in enclosures with diffuse noise and different speaker-microphone configurations. Instead, the performance in the moving vehicle environment is slightly lower with respect to the SOTA.

Prior works were not able to generalize when no samples of the target environment are present in the training set. Instead, our approach is able to better recognize the replay attack in all the environment with large EER improvements.

D. Analysis on fixed and learnable beamformer

To demonstrate the effectiveness of the adaptive learnable beamformer, Table 3 shows the performance of different configurations of the beamformer, following the same experimental setup as in Section B. We analyzed the scenarios

TABLE 2. EERs (%) for environment-dependent vs. environment-independent scenarios on ReMASC dataset.

Methods	Env-A	Env-B*	Env-C	Env-D
Environment-dependent				
CQCC-GMM [15]	13.5	17.4	21.3	22.1
ETECC [19]	15.1	33.8	14.1	10.4
CTECC [18]	13.0	26.8	9.9	10.1
TECC [17]	13.4	27.9	10.3	**9.1**
M-ALRAD	$\mathbf{8.1} \pm 2.3$	$\mathbf{5.8} \pm 2.4$	$\mathbf{7.5} \pm 2.6$	14.0 ± 2.0
Environment-independent				
CQCC-GMM [15]	19.9	39.9	34.6	48.9
ETECC [19]	29.0	32.5	30.0	49.9
CTECC [18]	28.5	34.7	32.5	50.0
TECC [17]	26.8	35.4	31.8	50.0
M-ALRAD	$\mathbf{13.8} \pm 2.8$	$\mathbf{20.3} \pm 3.1$	$\mathbf{15.1} \pm 5.5$	$\mathbf{24.2} \pm 2.4$

Env-B* does not encompass D1 genuine utterances due to hardware fault during data collection.

where (i) a learnable beamformer is present but not adaptive, i.e., single-channel weights $\mathbf{W} \in \mathbb{C}^{F}$ are learnable during training and fixed during inference, (ii) same as (i) but with multi-channel weights $\mathbf{W} \in \mathbb{C}^{N \times F}$, (iii) M-ALRAD without any regularization, (iv) M-ALRAD without orthogonality regularization, and (v) M-ALRAD without sparsity constraint.

It is important to note that the fixed beamformer, i.e., with predetermined weights $\mathbf{W}$ for all audio recordings, performs similarly to the variable version when the number of channels is low (2 and 4 channel microphones). However, this configuration performs worse with the arrays D3 and D4, with a decrease in performance of 6% and 1.3%. This could be due to the number of microphones in these arrays, where fixed complex beamforming weights cannot adapt to more general scenarios.

In fact, using the adaptive beamformer in which the weights $\mathbf{W}$ depend on the input audio, it is possible to better generalize the task, providing better performance in scenarios where SOTA models struggled to identify replay attacks due to varying spatial characteristics.

Similarly, beamformer regularizations are more effective with larger microphone arrays, as it can be inspected from Table 3. Moreover, the combination of orthogonality and sparsity constraints yielded the best performance, although small EER improvements have been achieved by using only one regularizer.

TABLE 3. Ablation study with respect to EER (%).

Methods	D1	D2	D3	D4
(i) Learnable fixed single-channel bm	7.3 ± 2.6	12.3 ± 2.5	14.6 ± 1.1	19.9 ± 3.1
(ii) Learnable fixed multi-channel bm	5.8 ± 1.3	11.4 ± 4.1	16.5 ± 1.5	15.5 ± 2.1
(iii) Without any regularization	7.3 ± 1.8	12.2 ± 1.7	11.2 ± 4.5	23.0 ± 5.2
(iv) Without orthogonality	6.2 ± 2.4	9.4 ± 1.6	10.6 ± 5.8	17.1 ± 2.9
(v) Without sparsity	6.6 ± 1.0	8.7 ± 2.2	11.2 ± 4.5	17.8 ± 4.5
M-ALRAD	$\mathbf{5.2} \pm 1.2$	$\mathbf{10.0} \pm 0.9$	$\mathbf{10.4} \pm 2.1$	$\mathbf{14.2} \pm 0.8$

E. Comparison with state-of-the-art

Since several previous works did not provide microphone-wise performance, we report the mean EER, namely mEER, across the entire test set in Table 4. From the global results, the superiority of our approach is clear. Although we were able to decrease the mEER from 14.7% to 10.0%, this error rate still indicates significant room for improvement with respect to other biometric traits.

TABLE 4. Comparison of mEER (%) with state-of-the-art approaches on evaluation subset of ReMASC [15].

Methods	mEER (%)
NN-Single [3]	22.9
NN-Dummy Multichannel [3]	22.2
Lexa [25]	18.9
TECC [26]	17.9
NN-Multichannel [3]	16.7
ETECC [19]	15.4
LCNN+GMM [18]	14.0
TECC [17]	14.7
M-ALRAD	**10.0**

It is important to highlight that other versions of M-ALRAD, such as those investigated in Section D, gave better performance from the SOTA, except for the case where a single-channel is replicated, i.e., without exploiting spatial information. This suggests that the combination of a learnable beamformer and a CRNN as a classifier is enabling SOTA performance.

IV. Conclusions

In this work, we proposed a novel replay attack detection, namely M-ALRAD, based on a end-to-end DNN, which includes a learnable and adaptive beamformer and a CRNN that exploits spatial cues from the beamformed signal. Specifically, it performs both time-frequency beamforming and detection from multi-channel recordings. Experimental results on the ReMASC dataset showed the superiority of our approach with respect to the SOTA, reducing the mEER by 4.7%. In addition, we designed a single-channel version, namely ALRAD, to prove the efficacy to exploit spatial cues in detecting replay speech attack. Finally, we demonstrated that the use of an adaptive time-frequency beamformer with multi-channel weights is helpful to tackle the replay speech attack both in seen and unseen scenarios, providing better generalization capabilities with respect to the SOTA. The analysis of microphone array mismatches, including different number of geometries and microphones, for replay speech detection using multi-channel recordings is set as future work.

REFERENCES

[1] W. Huang, W. Tang, H. Jiang, J. Luo, and Y. Zhang, "Stop deceiving! an effective defense scheme against voice impersonation attacks on smart devices," *IEEE Internet of Things Journal*, vol. 9, no. 7, pp. 5304–5314, 2022.

[2] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[3] Y. Gong, J. Yang, and C. Poellabauer, "Detecting replay attacks using multi-channel audio: A neural network-based method," *IEEE Signal Processing Letters*, vol. 27, pp. 920–924, 2020.

[4] B. Chettri, E. Benetos, and B. L. T. Sturm, "Dataset Artefacts in Anti-Spoofing Systems: A Case Study on the ASVspoof 2017 Benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020.

[5] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine*, 2004.

[6] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[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 *Interspeech*, 2017.

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech*, 2019.

[9] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A Capsule Network Based Approach for Detection of Audio Spoofing Attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[10] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," *Machine Learning With Applications*, vol. 14, p. 100503, 2023.

[11] L. Xu, J. Yang, C. H. You, X. Qian, and D. Huang, "Device Features Based on Linear Transformation With Parallel Training Data for Replay Speech Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1574–1586, 2023.

[12] R. He, Y. Cheng, Z. Zheng, X. Ji, and W. Xu, "Fast and Lightweight Voice Replay Attack Detection via Time-Frequency Spectrum Difference," *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 29 798–29 810, 2024.

[13] J. Xue, C. Fan, J. Yi, J. Zhou, and Z. Lv, "Dynamic Ensemble Teacher-Student Distillation Framework for Light-Weight Fake Audio Detection," *IEEE Signal Processing Letters*, vol. 31, pp. 2305–2309, 2024.

[14] M. Omologo, M. Matassoni, and P. Svaizer, "Speech recognition with microphone arrays," in *Microphone arrays: signal processing techniques and applications*. Springer, 2001, pp. 331–353.

[15] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," *Interspeech*, 2019.

[16] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1080–1091.

[17] H. Kotta, A. T. Patil, R. Acharya, and H. A. Patil, "Subband Channel Selection using TEO for Replay Spoof Detection in Voice Assistants," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020.

[18] R. Acharya, H. Kotta, A. T. Patil, and H. A. Patil, "Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6364–6368.

[19] A. T. Patil, R. Acharya, H. A. Patil, and R. C. Guido, "Improving the potential of enhanced teager energy cepstral coefficients (ETECC) for replay attack detection," *Computer Speech & Language*, vol. 72, p. 101281, 2022.

[20] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[21] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

<Society logo(s) and publication title will appear here.>

[22] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Speaker Distance Estimation in Enclosures From Single-Channel Audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2242–2254, 2024.

[23] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Single-Channel Speaker Distance Estimation in Reverberant Environments," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[24] C. Djork-Arné, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in *International Conference on Learning Representations (ICLR)*, 2016.

[25] R. Rodriguez and Y. Li, "Lexa: A Liveness Detection Enabled Voice Assistant," in *IEEE World AI IoT Congress (AIIoT)*, 2023.

[26] P. K. Chodingala, S. S. Chaturvedi, A. T. Patil, and H. A. Patil, "Robustness of das beamformer over mvdr for replay attack detection on voice assistants," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022.