Impact of Microphone Array Mismatches to Learning-based Replay Speech Detection

Michael Neri, Tuomas Virtanen
Faculty of Information Technology and Commmunication Sciences,
Tampere University, Tampere, Finland
{michael.neri, tuomas.virtanen}@tuni.fi

***Abstract*—In this work, we investigate the generalization of a multi-channel learning-based replay speech detector, which employs adaptive beamforming and detection, across different microphone arrays. In general, deep neural network-based microphone array processing techniques generalize poorly to unseen array types, i.e., showing a significant training-test mismatch of performance. We employ the ReMASC dataset to analyze performance degradation due to inter- and intra-device mismatches, assessing both single- and multi-channel configurations. Furthermore, we explore fine-tuning to mitigate the performance loss when transitioning to unseen microphone arrays. Our findings reveal that array mismatches significantly decrease detection accuracy, with intra-device generalization being more robust than inter-device. However, fine-tuning with as little as ten minutes of target data can effectively recover performance, providing insights for practical deployment of replay detection systems in heterogeneous automatic speaker verification environments.**

***Index Terms*—Replay attack, Physical Access, Beamforming, Spatial Audio, Voice anti-spoofing, Array Mismatch.**

I. Introduction

Recently, voice assistants (VAs) have been employed for human-machine interaction, exploiting voice as a biometric trait for authorizing the user [1]. In fact, using VAs we can control smart devices in our homes through Internet of things (IoT) networks and transmit sensitive information over the Internet. Malicious users can employ attacks on audio recordings [2] to access personal data and devices by deceiving the automatic speaker verification (ASV) of VAs. By applying text-to-speech (TTS) and/or voice conversion (VC) algorithms, it is possible to perform a logical access (LA) attack which modifies both talker traits or speech content. In the same direction, compressing or quantizing speech recordings can mask LA attacks by introducing artifacts, namely deepfake (DF) attack [2]. Physical access (PA) attacks focus on deceiving the ASV system at the microphone level [3]. An attacker can mimic user's speech to access to sensitive data or use a spoofing microphone to later replay the recorded speech to the ASV, which are called *impersonation attack* [1] and *replay attacks* [4] respectively. We focus on *replay attack* as, differently from other biometric markers, voice is easy to be captured by attackers to from the target speaker [5]. Moreover, ASV struggles to distinguish between genuine and replay speech even with commercial-off-the-shelf (COTS) devices like smartphones and loudspeakers [6], [7].

Prior works aimed to tackle PA attacks by collecting single-channel recordings, such as RedDots [3], ASVSpoof2017 PA [8], ASVSpoof2019 [9] PA, and ASVSpoof2021 PA [2]. From these datasets, several single-channel replay attack detection models have been designed, employing both deep neural network (DNN) [10], [11] and hand-crafted time-frequency features [12], [13]. However, single-channel replay detectors demonstrated poor generalization capabilities across different environments and devices, similarly to what happens in different tasks like in acoustic scene classification and unsupervised anomalous sound detection [14], [15]. To cope with this issue, previous works collected more diverse datasets [14], [15], encompassing different environments and microphones, or defining ad-hoc training procedures to minimize the domain shift [16].

For speech enhancement and separation tasks, microphone arrays are used in ASV systems to exploit spatial information and improve audio quality [17]. However, training-testing mismatches of microphone arrays is typically a severe issue in DNN-based microphone array processing, thus yielding poor generalization capabilities to unseen testing devices. To study the impact of these training-testing differences, in [18] the authors conducted initial studies on mismatched noise environments, different microphones, simulated, and real data for speech enhancement and separation. In [19] a fixed beamformer in conjuction with a recurrent neural network (RNN) has been analyzed in terms of speech enhancement performance changing the arrangement of microphones in an array. Similarly, Han *et al.* [20] designed an unsupervised method for adapting a single-channel DNN to seen array configurations. In the context of speaker diarization, a DNN has been designed to be robust to array mismatch in [21]. Recent works have explored to include array geometry information to DNN to improve robustness of learning-based approaches to unseen devices [22], [23].

Regarding the replay speech detection attack, the lack of datasets and models on this task makes the microphone array mismatch more challenging also with unseen factors such as environments [6] and different array geometries. To the best of our knowledge, studying the impact of microphone mismatches with diverse devices and number of microphones has never been explored in the ASV context.

To better study this phenomenon, we reply to the following research questions:

- *RQ1: How much is the degradation of the detection performance of a learning-based detector with microphone*

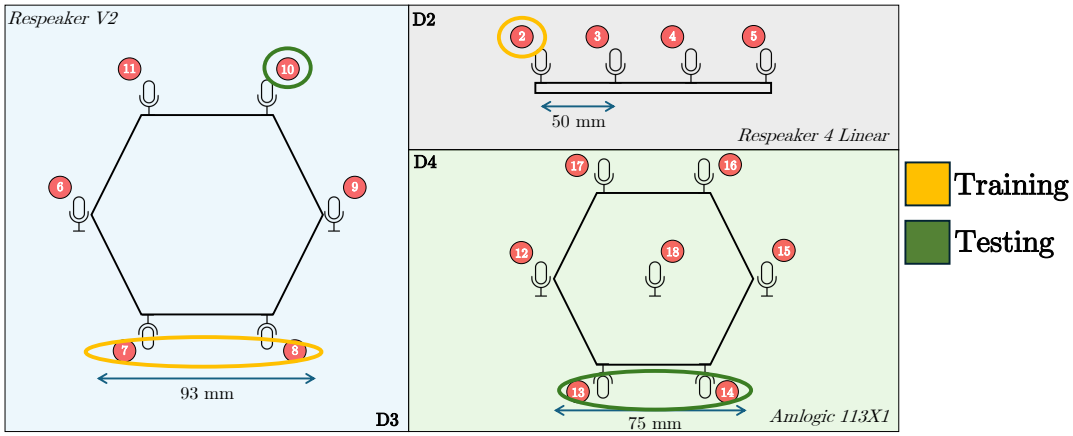


Fig. 1. Overview of microphone arrays from ReMASC [7] dataset involved in the experiments. Few examples of training and testing configurations are depicted. For instance, a model is trained on microphone with ID 2 from D2 and tested on microphone ID 10 from D3 with speech recordings in the same environment. An example of multi-channel setup is training a model on the couple $(7, 8)$ from D3 and testing on the couple $(13, 14)$ from D4.

array mismatch?

- *RQ2: Is fine-tuning a learning-based detector trained on device* D_i *to target device* D_j *beneficial?*
- *RQ3: How much data is required to recover the lower-bound equal error rate (EER) of the target array from a pre-trained model on a source array (assuming that the lower-bound is a model trained solely on target data)?*

To answer these research questions, we employ Realistic Replay Attack Microphone Array Speech (ReMASC) dataset [7], which is the only replay speech dataset that encompasses real recordings from different microphone arrays. Moreover, we employ the state-of-the-art replay detector M-ALRAD [6], which combines adaptive beamforming and detection, achieving the best detection performance.

The structure of the paper is organized as follows. Section II depicts the model and the dataset used for evaluation of microphone array mismatch for replay speech detection. Section III shows the results of the performed analysis, changing the number of microphones involved in an array and varying the training and the testing setup. Finally, Section IV draws the conclusions and future directions to mitigate the problem.

II. Materials

This section describes the ReMASC dataset and the model M-ALRAD (Multi-channel Replay Attack Detector) employed in the evaluation of performance in mismatched microphone arrays.

A. Dataset under analysis

ReMASC [7] has been selected since it is the only dataset providing multi-channel recordings, which have been collected by devices with different number of microphones and geometries, for this task. All recordings are synchronized across four parallel ASV microphone arrays $\mathrm{D}1, \mathrm{D}2, \mathrm{D}3, \mathrm{D}4$ with 2, 4, 6, and 7 omnidirectional microphones, respectively, with 9240 genuine and 45472 replay audio samples. Moreover, D4's sampling rate is 16 kHz, differently from D1, D2, and D3 which operate at 44.1 kHz. Bit depth is 16 bits for all devices, except for D3 which is 32 bits. Overall, ReMASC features 55 diverse speakers, varying in gender and vocal characteristics, and includes recordings from four distinct environments: an outdoor setting (Env-A), two enclosed spaces (Env-B and Env-C), and a moving vehicle (Env-D), capturing different acoustic conditions. The dataset employs two spoofing microphones to record genuine speech, which is then replayed through four playback devices of varying quality. Figure 1 illustrates the structure of microphone arrays and the identification code for each single-channel omnidirectional microphone. To select a microphone from the dataset, we label each microphone with an ID from 2 to 18.

B. Selected replay speech detector

In our experiments, we employ M-ALRAD [6], which is a convolutional recurrent neural network (CRNN)-based approach that jointly processes N multi-channel complex short-time Fourier transforms (STFTs) $\{X_{\mathrm{STFT}_{n,T,F}}, n = 1, \ldots, N\}$ with T and F time and frequency bins, respectively, to produce a single-channel beamformed spectrogram for detecting replay speeches. First, a convolutional neural network (CNN) $f_{BM} : \mathbb{C}^{N\times T\times F} \rightarrow \mathbb{C}^{T\times F}$ outputs beamforming weights $\mathbf{W} \in \mathbb{C}^{N\times T\times F}$ through a series of Conv2D-BatchNorm-ELU-Conv2D operations, where real and imaginary parts are concatenated along the channel dimension. The beamformed spectrogram $\hat{X}_{\mathrm{STFT}}$ is computed as

$$\hat{X}_{\mathrm{STFT}_{T,F}} = \sum_{n} X_{\mathrm{STFT}_{n,T,F}} \cdot \mathbf{W}_{n,T,F}. \quad (1)$$

The beamformed spectrogram is then processed by a CRNN classifier, previously used for speaker distance estimation [24], [25]. Magnitude and phase features of the beamformed STFT

are extracted, with sin&cos transformations applied to the phase. These features are arranged in a $T \times F \times 3$ tensor and passed through three convolutional layers with 1×3 filters and batch normalization, followed by parallel max and average pooling. Two bi-directional gated recurrent unit (GRU) layers with 128 neurons refine the feature maps, and the final hidden state is mapped to the binary prediction $\hat{y} \in \mathbb{R}$ exploiting a fully connected layer with softmax activation function, where 0 denotes the genuine class and 1 represents the replay class.

The model is trained to minimize the binary cross-entropy loss between predicted and true binary classes, i.e., genuine or replay. In addition, orthogonality and sparsity losses are employed using the same hyperparameters as in [6].

III. Experimental results

In this section, we evaluate the detection performance of M-ALRAD when different microphone configurations are used in the training and testing sets. First, details on the analysis regarding the subset of ReMASC and training hyperparameters are provided. Then, we first evaluate a single-channel mismatch scenario. Next, we increase the number of microphones from two to four, in order to reply to RQ1. Finally, we perform fine-tuning to provide some insights to RQ2 and RQ3.

A. Metric and implementation details

We employed recordings from Env-B where the configuration of talker/loudspeaker and microphone are varied in a grid whereas the room is the same. The design rationale of this choice is to focus on microphone array mismatches without the influence of varying the environment. The selected microphone arrays from ReMASC are $\{\mathrm{D2}, \mathrm{D3}, \mathrm{D4}\}$ since D1 does not have genuine speech recordings due to hardware fault during data collection [7].

Following [6], each model is trained or fine-tuned for each setup of microphones with a batch size of 32 using a cosine annealing scheduled learning rate of 0.001 for 50 epochs. We only analyze the first second of multi-channel recordings to reduce the computational complexity of the approach [4], [6]. The STFT is computed with a Hanning window of length 32 and 46 ms with 50% overlap for 44.1 and 16 kHz, respectively. When sampling rate are different, e.g., training on D2 and testing on D4 and viceversa, upsampling or downsampling is performed on the target recording in order to use the same STFT parameters. To evaluate the replay speech attack detection, the EER metric is used following the same train-test split provided by the authors of the ReMASC dataset [7]. Five independent runs are collected to compute the 95% mean confidence intervals. Overlapping training and testing channels are not performed, e.g., training on $(2, 3)$ and testing on $(3, 4)$, for avoiding data leakage.

B. Single-channel mismatch

First, we evaluate the performance of training a model on a single microphone and testing on a different one. By doing so, it is possible to jointly assess (i) *inter-device* mismatches, e.g., training on microphone ID 2 and testing on microphone ID 4 of array D2, and (ii) *intra-device* mismatches, e.g., training on microphone with ID 1 from array D2 and testing on microphone with ID 6 from array D3. All the results have been collected in a matrix visualization in Figure 2, where rows denote the training microphone ID and columns are the testing microphone ID. It is clear from the results that *inter-device* experiments show better performance, as microphones belonging to the same array share similar acoustic characteristics like frequency responses, as previously demonstrated in prior works [18]. Instead, *intra-device* results are worse than random guess (i.e., EER $\geq$ 50%). In addition, microphones belonging to D3 on average performs worse than D2 and D4 both in the *intra-* and *inter-device* scenarios. To analyze this behavior, the average STFT magnitude of each microphone in array D3 is depicted in Figure 3, which shows significant differences between microphones, depicting dominant spectral peaks which are different for different microphones. As the model is trained using features derive from the STFT, these difference cause domain shift in *intra-device* results.

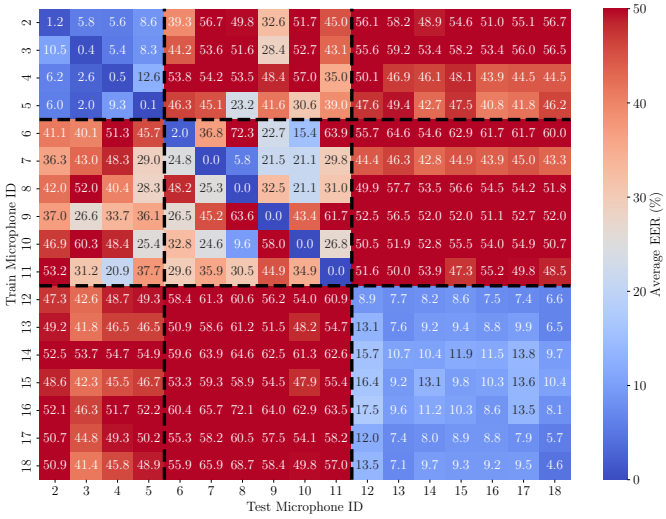


Fig. 2. Single-channel cross training-testing EERs.

C. Multi-channel mismatch

With two-channel setups in Figure 4, the overall error rates in matched conditions decrease compared to single channel, indicating that combining two microphone inputs helps in improving the detection performance. Logically, *intra-device* pairs continue to show the best results. When switching to significantly different pairs, such as $(6, 9)$ to $(12, 15)$, the degradation is more evident, reinforcing that spatial and array-level shifts pose a difficult generalization challenge.

The three-channel matrix in Figure 5 shows a similar trend with respect to the two channel configuration. In fact, for instance training on $(2, 3, 4)$ from D2 and testing on $(12, 17, 15)$ from D4, still result in performance drops and random guess predictions.

The four-channel matrix in Figure 6 reveals that *intra-device* configurations continue to achieve the lowest error

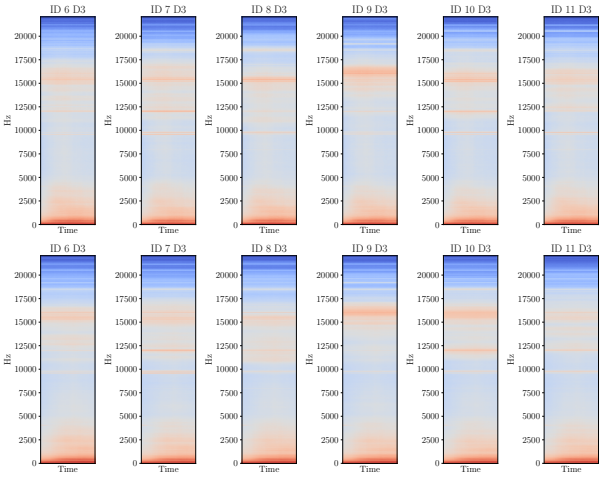


Fig. 3. Average STFT magnitudes from D3 in Env-B. The first row consists of genuine speeches whereas the second row consists of replay speeches.

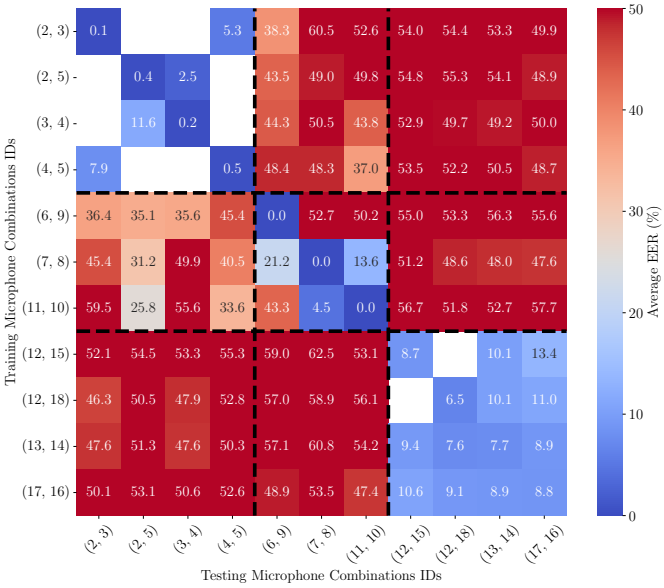


Fig. 4. Two channels cross training-testing EERs.

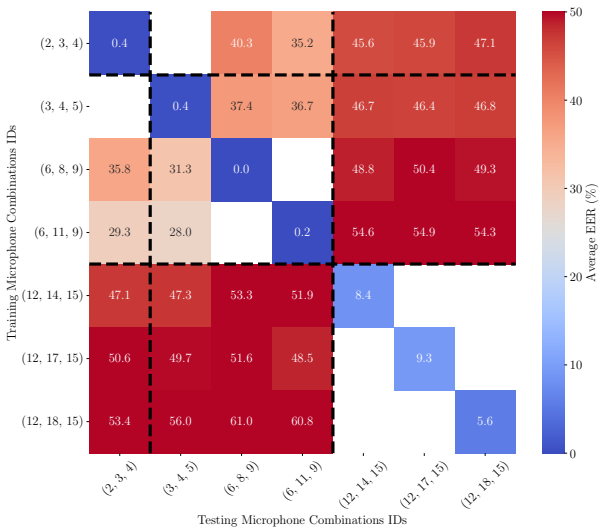


Fig. 5. Three channels cross training-testing EERs.

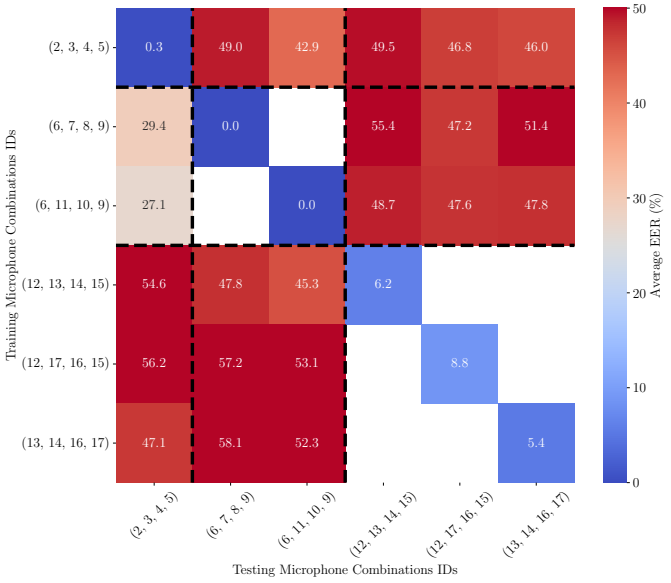


Fig. 6. Four channels cross training-testing EERs.

rates. However, non-random guess performance are achieved by training M-ALRAD on four microphones from D2 and testing on D3. Clearly, generalizing on the array D4 is more challenging since this device operates at a different sample rate.

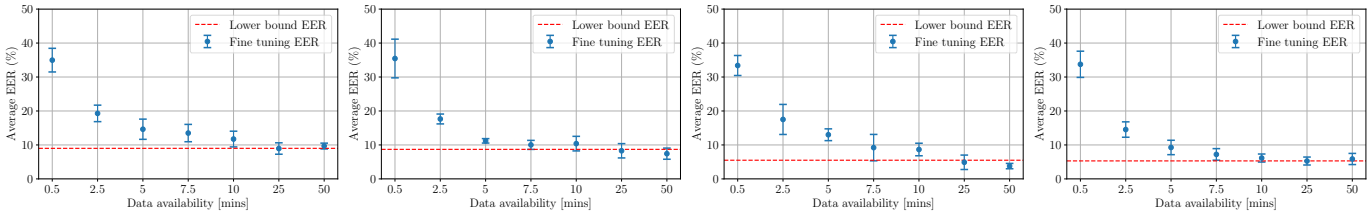
D. Fine-tuning with target data

Addressing RQ2 and RQ3, we fine-tune a pretrained model on a source array D_i to an unseen target array D_j, with various number of microphones. To study this adaptation, fine-tuning was performed by varying the amount of data available from target domain from half a minute to fifty minutes. Analyzing this behavior is particularly important in real-world applications where collecting large amounts of data may not always be feasible [19]. In fact, from this study is possible to determine how much data is practically needed to achieve a desired level of model's performance.

Figure 7 shows the EERs of the fine-tuned M-ALRAD for one microphone (7a), two microphones (7b), three microphones (7c), and four microphones (7d). It is worth highlighting that in all scenarios the EER lower bound, e.g., a version of the model trained solely on target data, is reached by using at least ten minutes of target recordings.

IV. Conclusion

This study analyzed the microphone array mismatch problem in replay speech detection. Using a state-of-the-art learning-based detector, which also performs adaptive beamforming, we demonstrated that single-channel recordings suffers from generalization across different devices, while multi-



(a) Source ID 2 and target ID 12. (b) Source IDs $(2, 3)$ and target IDs $(12, 15)$. (c) Source IDs $(2, 3, 4)$ and target IDs $(12, 18, 15)$. (d) Source IDs $(2, 3, 4, 5)$ and target IDs $(12, 14, 16, 17)$.

Fig. 7. Confidence interval of EER of M-ALRAD by fine-tuning the model with different availabilities of training data for (a) one, (b) two, (c) three, and (d) four microphones.

channel configurations slightly improve the detection rate but remain susceptible to cross-array variability. If there is an availability of target array recordings, it is possible to adapt a pre-trained model by fine-tuning, showing that even a limited amount of target-domain data can substantially restore detection performance. Future research should explore array geometry-aware learning approaches, thus including some geometry features or a-priori knowledge of the microphone array, and domain adaptation techniques to further enhance robustness, i.e., generalization capabilities, against microphone mismatches.

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] 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.

[4] 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.

[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] M. Neri and T. Virtanen, "Multi-channel replay speech detection using an adaptive learnable beamformer," *IEEE Open Journal of Signal Processing*, pp. 1–7, 2025.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *DCASE*, 2020.

[15] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and Discussion on DCASE 2022 Challenge Task 2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques," in *DCASE*, 2022.

[16] M. Neri and M. Carli, "Semi-Supervised Acoustic Scene Classification Under Domain Shift Using an Attention Module and Angular Loss," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2024.

[17] 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.

[18] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[19] J. Lin, N. Moritz, Y. Huang, R. Xie, M. Sun, C. Fuegen, and F. Seide, "AGADIR: Towards Array-Geometry Agnostic Directional Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[20] C. Han, K. Wilson, S. Wisdom, and J. R. Hershey, "Unsupervised Multi-Channel Separation And Adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[21] T. Mariotte, A. Larcher, S. Montrésor, and J. Thomas, "Channel-Combination Algorithms for Robust Distant Voice Activity and Overlapped Speech Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1859–1872, 2024.

[22] M. Heikkinen, A. Politis, and T. Virtanen, "Neural Ambisonics Encoding For Compact Irregular Microphone Arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[23] C. Pan, J. Chen, and J. Benesty, "An approach to microphone array geometry transform," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1–14, 2025.

[24] 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.

[25] 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.