Information Fusion 105 (2024) 102257

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection

Rui Liu *, Jinhua Zhang, Guanglai Gao

Department of Computer Science, Inner Mongolia University, Hohhot, 010021, China

ARTICLE INFO

Keywords:
Audio deepfake detection (ADD)
Mono-to-binaural conversion
Multi-space channel representation (MSCR) learning

ABSTRACT

Audio deepfake detection (ADD) aims to detect the fake audio generated by text-to-speech (TTS), and voice conversion (VC), etc., which is an emerging topic. Traditionally we read the mono signal and analyze the artifacts directly. Recently, the mono-to-binaural conversion based ADD approach has attracted increasing attention since the binaural audio signals provide a unique and comprehensive perspective on speech perception. Such method attempts tried to first convert the mono audio into binaural, then process the left and right channels respectively to discover authenticity cues. However, the acoustic information from the two channels exhibits both differences and similarities, which have not been thoroughly explored in previous research. To address this issue, we propose a new mono-to-binaural conversion based ADD framework that considers multi-space channel representation learning, termed "MSCR-ADD". Specifically, (1) the feature representations of the respective channels are learned by the channel-specific encoder and stored in the *channel-specific space*; (2) the feature representations capturing the difference between the two channels are learned by the channel-differential encoder and stored in the *channel-differential space*; (3) after which the channel-invariant encoder learn the channel commonality representations in the *channel-invariant space*. Note that we propose orthogonal and mutual information maximization losses to constrain the channel-specific and invariant encoders. At last, three representations from various spaces are mixed together to finalize the deepfake detection. It is worth noting that the feature representations in the channel-differential and invariant spaces unveil the differences and similarities between the two channels in binaural audio, enabling us to effectively detect artifacts in fake audio. The experimental results on four benchmark datasets demonstrate that our MSCR-ADD is superior to existing state-of-the-art approaches.

1. Introduction

Audio Deepfake Detection (ADD) technology is a task that determines whether the given audio is authentic or counterfeited [1–3], and it has attracted increasing attention recently. Specifically, with the rapid progress of Text-To-Speech (TTS) synthesis [4–9] and Voice Conversion (VC) techniques [10,11], one is able to impersonate another's voice easily. However, such technology can be abused. More and more deepfake audio has been used for malicious purposes, such as in the spreading of rumors and fraud cases. This calls for the effective detection of fake audio [12].

Note that conventional methods for ADD mainly focus on two directions, including (1) robust feature extraction [13,14] and (2) effective model design [15–17]. These methods have achieved remarkable performance using mono audio as input and modeling speech forgery cues in it. Currently, mono-to-binaural conversion based ADD approaches are gradually coming to the forefront. Unlike the mono signal just performs single-channel, binaural audio includes dual-channel, that are left and right-channel signals, and provides a unique and comprehensive perspective on speech perception. To this end, existing methods mainly tried to read and process the left and right channel signals respectively, and then fuse them together to finalize the deepfake detection. For example, M2S-ADD [18] employed a graph attention-based dual-branch neural architecture to extract the deep features of the left and right channel signals. Finally, the two styles of information are also fused by graph attention to make the final decision. By incorporating two-channel binaural audio information compared to mono speech, the fake audio could reveal itself and be successfully detected.

Despite the advances, the signals of the left and right channels have clearly behaved differently, and similarities in binaural audio [19], which has not been studied in the prior work. Specifically, we visualize the spectral feature, that is linear frequency cepstral coefficients (LFCC) in this work, for bonafide ad fake audios in terms of mono and binaural formats in Fig. 1. White and blue boxes are used to highlight the

* Corresponding author.
E-mail address: liurui_imu@163.com (R. Liu).

https://doi.org/10.1016/j.inffus.2024.102257
Received 9 November 2023; Received in revised form 10 January 2024; Accepted 11 January 2024
Available online 14 January 2024

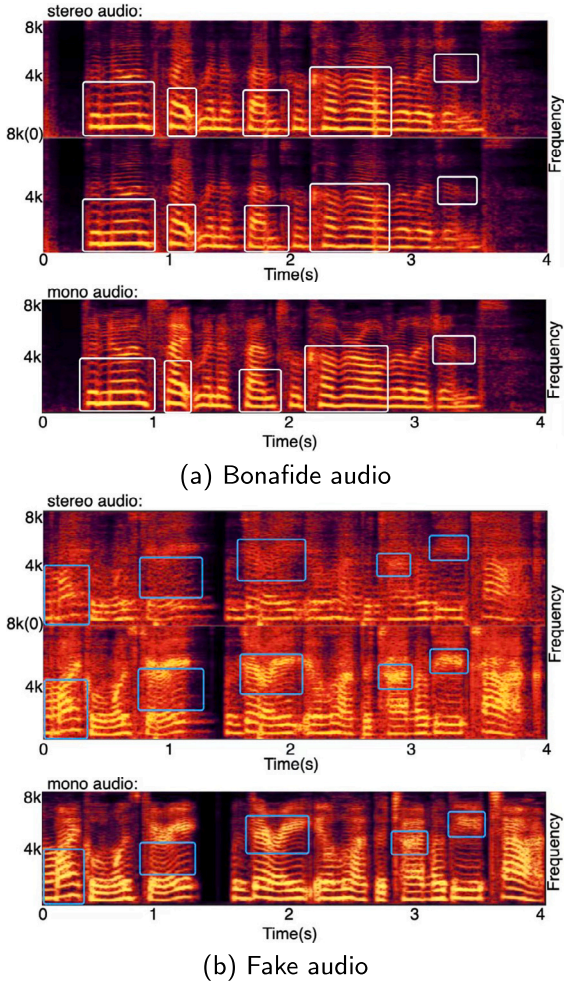

Fig. 1. Comparison of visualization analysis in the spectral domain. (a) shows the spectral details of the mono and binaural signals for bonafide audio; (b) shows the spectral details for fake audio. Unlike the white boxes, the blue boxes show that the spectral artifacts are particularly exposed when the mono fake audio is converted to binaural audio. It indicates that the left and right channels for fake audio differ markedly in spectral details, in addition to the natural similarities in spectral structure.

spectral details at specific locations for bonafide and fake audios respectively. As can be seen from the blue boxes for fake audio, the original spectral information is not fully reflected, and there is even a lot of noise, in the left and right channels of the binaural panel after mono-to-binaural conversion. Note that all the spectral details for bonafide audio, as in the white boxes, are preserved. It indicates that the left and right channels for fake audio differ markedly in spectral details, in addition to the natural similarities in spectral structure. In addition, in the mono-to-binaural generation field, Leng et al. [19] decomposed the synthesis of binaural audio into two steps. The common information of the two channels of binaural audio is generated in the first stage, and their difference is generated in the second stage. However, existing mono-to-binaural conversion based ADD works usually fail to exploit such differences and similarities between two channels, thus limiting their performance. Inspired by the above observations, can we fully explore the differences and commonalities between the two channels for more accurate ADD, which is the focus of this work.

In this paper, we propose a new mono-to-binaural conversion based ADD framework that considers Multi-Space Channel Representation learning, termed "MSCR-ADD". Specifically, MSCR-ADD learns three distinct frame-level feature representations for each channel in three various spaces, including channel-specific, invariant, and differential

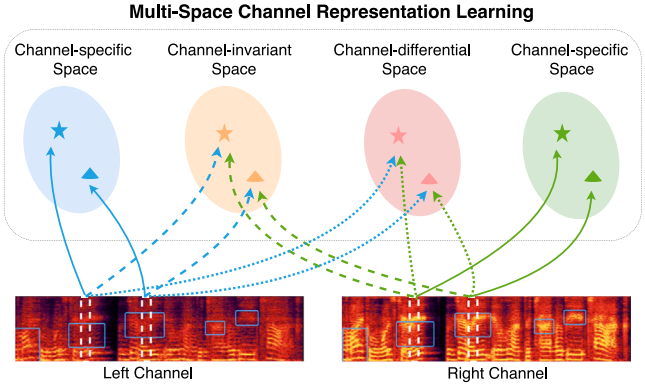

Fig. 2. Learning multi-space channel representations for binaural audio-based ADD through channel-specific, differential, and invariant spaces. These frame-level features are later utilized for fusion and final audio real/fake detection.

spaces. As shown in Fig. 2, (1) Channel-specific space seeks to learn the specific acoustic features for the left and right channels respectively, that are private to each channel. For any dual-channel binaural audio, each channel holds distinctive characteristics due to the differences in the information received by the left and right ears of the human brain. To this end, we adopt an orthogonal [20] based channel-specific encoder to generate the channel-specific acoustic feature sequence for two channels. (2) Channel-invariant space attempts to map two channels to a shared feature space with frame-level acoustic consistency. We design a mutual information maximization based channel-invariant encoder to learn the channel commonality representations between two channels. Note that mutual information is used to further enrich the acoustic perception information for the channel invariant representation. (3) Channel-differential space aims to learn the differences in the spectral domain between the two channels. We first calculate the difference between the two spectra from two channels and then obtain the high-level channel-differential feature representation using the channel-differential encoder. The prior work does not utilize such differential before fusion, which puts a limit on forgery clue discovery. Learning such modality-differential and invariant features, thus, complements the specific latent features captured in the specific space and provides a comprehensive feature representation of the left and right channels for binaural audio. Empirical results on four benchmark datasets demonstrate the effectiveness of our approach. The main contribution of this paper can be summarized as follows:

- Existing mono-to-binaural conversion based ADD works ignore the differences and commonalities between the two channels of binaural audio. We pioneer the definition of a feature representation of binaural audio as three spaces, including channel-specific, differential, and invariant spaces.
- We present MSCR-ADD, a new mono-to-binaural conversion based ADD scheme with a novel multi-space channel representation learning strategy. In this way, we provide a comprehensive feature representation for binaural audio, thus enhancing the ADD performance.
- Experimental results on four ASVspoof2019, ASVspoof2021, VSDC and PartialSpoof datasets verify the power of our method. MSCR-ADD is superior to the existing state-of-the-art.

The remainder of this paper is organized as follows: In Section 2, we briefly review some related works. In Section 3, we propose a novel MSCR-ADD framework. In Section 4, we introduce experimental datasets and setup in detail. In Section 5, we conduct experiments to verify the effectiveness of our method. Finally, we conclude this paper and discuss future work in Section 6.

2. Related works

2.1. Audio deepfake detection

With the rapid progress of speech generation, such as Text-To-Speech (TTS) [7,21] synthesis and Voice Conversion (VC) techniques [22], ADD technology aims to determine whether the given audio is authentic or counterfeited. Currently, the mainstream TTS and VC systems produce synthesized audio with a mono signal. Therefore, the ADD systems focus on how to process the mono audio and discover the fake cues. The conventional ADD approaches can be summarized into two categories: (1) robust feature extraction and (2) effective model design. For example, Patel et al. [23] propose to extract frame-level acoustic features based on cochlear filter spectrum coefficients and instantaneous frequency changes. Gupta et al. [14] proposed a novel Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA) features, with excellent temporal resolution as well as relative phase information. Some recent works [13,24] ingest raw audio [13] instead of hand-crafted acoustic features and achieve robust ADD performance. For model design, researchers adopt Gaussian mixture model [15], convolution neural network [25], deep neural networks [26], recurrent neural networks [16] and graph neural network [17] etc. to build the ADD architecture. All of the above work dealt with single-channel (mono) audio and showed that ADD is clearly achievable.

We note that two types of audio signals, namely mono and dual-channel binaural audio, can be converted to each other [27]. The binaural signals provide a unique perspective on the speech quality of the audio signal [28]. To this end, our preliminary work [18], termed M2S-ADD, attempts to discover audio authenticity cues during the mono-to-binaural conversion process. This work also follows this spirit, however, there are some distinctive differences: (1) [18] dealt directly with time-domain signals, however, this work deals with spectral features, which have more acoustic details; (2) [18] first extracts the features of the left and right channels separately after using the graph attention to get the fusion features and then feed them into the classifier, while in this work, we design a novel multi-space channel representation learning to learn a more comprehensive feature representation for binaural audio. We also note that some ADD work [29] focuses on the fake binaural audio detection directly. However, it is a totally different task that our method tried to mine the authenticity cues from mono signals directly.

2.2. Binaural audio processing

Binaural audio processing has attracted increasing attention in recent years. There are many related fields, such as speech enhancement [28,30] and speech synthesis [19,31,32], to conduct research on binaural signal. For example, Delgado et al. [28] introduced a feature extracted from stereo audio signals aiming to represent a measure of perceived quality degradation in processed spatial auditory scenes. Tolooshams et al. [30] proposed a novel stereo-perception framework for speech enhancement. Compared with the traditional method with mono audio, using stereo can better preserve spatial images and enhance stereo mixture. Parida et al. [31] proposed an end-to-end trainable multi-modal transformer network with hierarchical multi-modal attention, for mono to binaural audio generation. From the observations of the above works, we can find that binaural audio also plays an important role in speech or audio information processing.

Therefore, whether we can improve ADD performance by incorporating binaural information is the main concern of our work. However, unlike the above works, this paper focuses on learning a comprehensive feature representation for the binaural signals.

2.3. Multi-space representation learning

Recent studies in many multimodal tasks suggest that the model benefits from multi-space representation learning including multimodal sentiment analysis [33–35], person re-identification [36–38], cross-modal retrieval [39] and image-sentence matching [40], etc. For example, MISA [33] maps the multimodal features into two spaces as modality-invariant and -specific representations, and then fuses them for downstream classification. MCLNet [38] learns modality-invariant representations by minimizing inter-modal discrepancy and maximizing cross-modal similarity. VI-REID [41] builds an individual network for each modality, with a shared identity loss to learn modality invariant representations.

However, these methods just map representations to two spaces, including modality-invariant or -specific spaces via similarity cost functions, while mono-to-binaural conversion based ADD tasks require frame-level acoustic differences and similarities from the left and right channels of binaural audio. To this end, we define three feature spaces, including channel-specific, differential, and invariant spaces, to provide a comprehensive frame-level feature representation of dual-channel signals.

3. Methodology

3.1. Overview

We propose a neural architecture, terms as MSCR-ADD, as shown in Fig. 1(a), that consists of a pre-trained mono-to-binaural (M2B) converter, an orthogonal based channel-specific encoder, a mutual information based channel-invariant encoder, a channel-differential encoder, a dynamic feature aggregator and an ADD classifier. Given the input time-domain mono signal X_m, the pre-trained M2B converter reads X_m and generates the binaural signal $\{X_l, X_r\}$ where X_l and X_r denote the left and right signals respectively. After that, feature sequence $\{X_l^{spec}, X_r^{spec}\}$ of spectral domain are extracted as our research object. To this end, these two sequences are then fed by channel-specific, invariant, and differential encoders respectively to generate channel-specific feature f_l^{spc}, f_r^{spc}, and channel-invariant, and differential features, i.e., $f_{lr}^{inv}, f_{lr}^{dif}$. The orthogonal based channel-specific encoder first considers the interaction of the left and right channels, then adopts orthogonal constraint to orthogonalize the f_l^{spc} and f_r^{spc}, thus obtaining the relatively independent channel-specific feature representations. Furthermore, the channel-invariant encoder first filters the channel-shared information from the channel-specific features, and then fuses the shared information together to form the channel-invariant features. The mutual information maximization strategy among the f_{lr}^{inv} and f_l^{spc}, f_r^{spc} is used to strengthen the channel agnosticism of f_{lr}^{inv}. Finally, all the above features from multi-space are fused with the multi-space feature aggregator module to adjust the contributions automatically according to the natural contribution differences. ADD classifier takes the joint feature to make the final decisions of bonafide or fake audio. In the following subsections, we will introduce the details of the multi-space channel representation learning.

3.2. Channel-specific representation learning

Channel-Specific representation learning is achieved by the Channel-Specific Encoder (CSE). As illustrated in Fig. 3(b), the channel-specific encoder employs a dual-branch architecture to learn channel-specific features f_l^{spc}, f_r^{spc}. First, we employ a pre-trained Whisper encoder [42] to extract the high-level acoustic features Y_l and Y_r for two channels. The key structure of Whisper is an off-the-shelf encoder–decoder Transformer architecture [43]. Its encoder is based on two convolutional layers, each processed by a GeLU activation function [44]. The information is later modified by adding position embeddings [43]. The

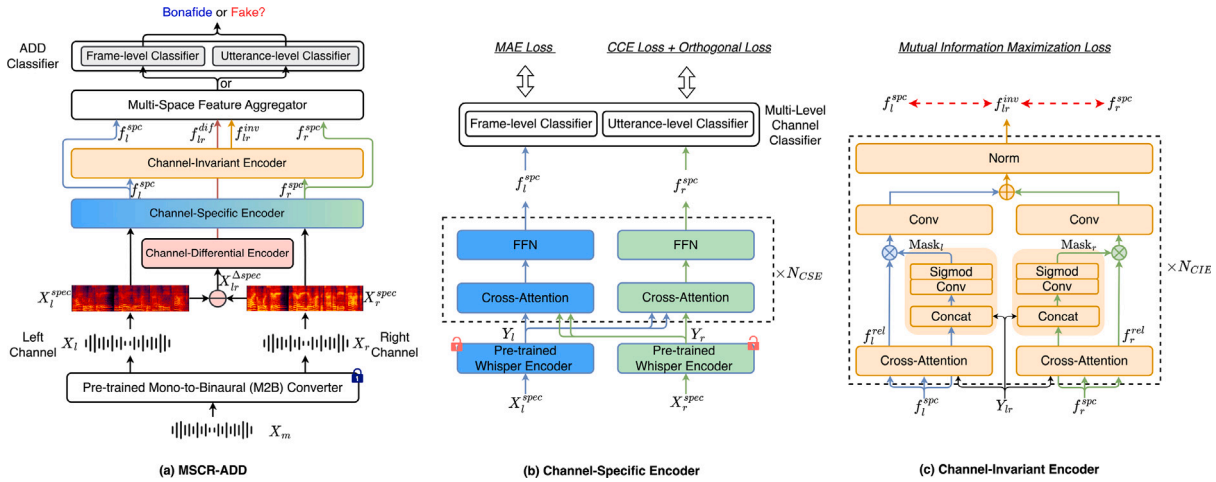

Fig. 3. Illustration of our MSCR-ADD. (a) The overall architecture of MSCR-ADD. (b) *Channel-Specific Encoder* that learns channel-specific representation f_l^{spc} and f_r^{spc}; (c) *Channel-Invariant Encoder* that strengthens the channel agnosticism of f_{lr}^{inv}.

encoder ends with a series of the pre-activation residual self-attention blocks [45], followed by the normalization layer. Such architecture allows Y_l and Y_r to capture the contextual dependencies within each channel. Note that Whisper [42] is a state-of-the-art automatic speech recognition (ASR) system. To fit the task of ADD, we unfreeze the Whisper encoder and update its parameters simultaneously with the training of the whole MSCR-ADD model. In this way, the Whisper encoder serves as the ADD-aware feature extractor rather than ASR.

$$Y_m = \text{Whisper}(X_m), m \in \{l, r\} \tag{1}$$

where the subscript m denotes channel index.

It is important to note that there are similar models to Whisper, such as Hubert [46] and Wav2vec [47] etc. These models also utilize transformer or CNN-based structures to learn advanced representations of speech signals, and these representations have been proven to play an important role in ADD tasks [48,49]. However, the input format for speech in Hubert and Wav2vec is in the time domain signals [46,47], while our model MSCR-ADD, similar to Whisper [42], uses frequency domain signals as input to explore the acoustic information in spectral parameters. In future work, it could be considered to supplement the existing features with features extracted from time-domain signals using models like Hubert or Wav2vec.

The following cross-attention mechanism tries to take the Y_l and Y_r and learn the frame-level interaction between the left and right channels. To take the left branch as an example, the extracted feature Y_r is treated as the Query input and Y_l is the Key and Value vector. The right branch is the opposite of this. Finally, there are feed-forward networks (FFN) to generate the channel-specific outputs. The cross-attention and FFN blocks are stacked by N_{CSE} layers to obtain the final channel-specific features f_l^{spc} and f_r^{spc}.

$$\begin{aligned} f_l^{spc} &= \text{FFN}(\text{Cross-Attention}(Y_r, Y_l, Y_l)) \\ f_r^{spc} &= \text{FFN}(\text{Cross-Attention}(Y_l, Y_r, Y_r)) \end{aligned} \tag{2}$$

To further enhance the frame-level discriminability of the channel-specific features f_l^{spc} and f_r^{spc}, we design a multi-level channel classifier that consists of frame-level and utterance-level classifiers.

- For the frame-level classifier, we sample half of the speech frames from each of the two channels in the same speech to form a mixed speech signal, after which we use BiLSTM and two fully connected (FC) layers for binary classification to determine whether each frame belongs to the left or right channel. The ground-truth category label is the real channel source of each speech frame. Here, we adopt Mean Absolute Error (MAE) loss $\mathcal{L}_{MAE}$ to optimize this classifier;
- For the utterance-level classifier, we employ BiLSTM and one FC layer to binary classify the current audio to determine whether the current acoustic signal belongs to the left or right channel. Again, the ground truth category label is the original channel information of each speech signal. We use Categorical Cross-Entropy (CCE) loss $\mathcal{L}_{CCE}$ to optimize this classifier. More importantly, we adopt the orthogonal loss $\mathcal{L}_{Ort}$ [20] to explicitly enhance the independence of f_l^{spc} and f_r^{spc}. Conventional methods are primarily devoted to keeping two channel-specific features far apart. However, it is hard to define the distance between them explicitly. Unlike the previous methods, we assume that two specific features under ideal conditions should be orthogonal to each other in two-dimensional space so that the relationship between the two can be more clearly constrained. Specifically, the objective of $\mathcal{L}_{Ort}$ is to enforce constraints to embedding space such that the embedding f_l^{spc} is orthogonal to the f_r^{spc}, which can effectively disentangle the channel-specific characteristics of each sample [50], further improving the discrimination of f_l^{spc} and f_r^{spc}. The objective function $\mathcal{L}_{Ort}$ is defined as:

$$\mathcal{L}_{Ort} = \sum \| f_l^{spc\top}, f_r^{spc} \|_F^2 \tag{3}$$

where $\| \cdot \|_F$ means the squared Frobenius norm [50].

3.3. Channel-invariant representation learning

After the channel-specific representation learning, the channel-invariant representation learning is achieved by the Channel-Invariant Encoder (CIE). Based on the learned channel-specific features f_l^{spc} and f_r^{spc}, the channel-invariant encoder first filters out the relevant information from the left and right channel-specific features with respect to the original left- and right-channel mixing information Y_{lr}, after which the filtered relevant information is fused together as the final channel-invariant feature f_{lr}^{inv}. On the top of the channel-invariant encoder, we design a mutual information maximization strategy to strengthen the channel agnosticism for the channel-invariant feature f_{lr}^{inv}.

As shown in Fig. 3(c), the channel-invariant encoder first involves a cross-attention module to extract the relevant information, f_l^{rel} and f_r^{rel}, for each channel-specific representation that is related to the mixing representation $Y_{lr} = \text{Add}(Y_l, Y_r)$, where Add means the element-wise addition operation. Note that the Y_{lr} contains the original acoustic feature of the fine-tuned Whisper encoder. Therefore, using its rich informativeness of Y_{lr} as a frame-level query can be good for relevant information extraction.

$$f_m^{rel} = \text{Cross-Attention}(Y_{lr}, f_m^{spc}, f_m^{spc}), m \in \{l, r\} \tag{4}$$

To further make the extracted feature invariant to two channels, we follow [51] and design a parallel convolutional network to learn a mask for filtering out the channel invariant information:

$$\text{Mask}_m = \sigma\left(\text{Conv}\left(f_m^{spe} \parallel Y_{lr}\right)\right), \tag{5}$$

where "Conv" denotes 1×1 convolutional layer, "$\parallel$" denotes feature concatenation, σ denotes Sigmoid activation.

After that, the relevant features of the left and right channels are multiplied by the learned mask to obtain the final channel-invariant features using another convolutional network and Norm operation.

$$f_{lr}^{inv} = \text{Norm}(\sum_{m\in\{l,r\}} \text{Conv}(f_m^{rel} \otimes \text{Mask}_m)) \tag{6}$$

where the "Norm" denotes layer normalization, "Conv" denotes 1×1 convolution followed by PReLU activation, $\otimes$ denotes element-wise multiplication.

To further make the final output f_{lr}^{inv} invariant to channels. we propose a new mutual information maximization loss $\mathcal{L}_{MIM}$ for the learned f_{lr}^{inv} and the channel-specific features f_l^{spc} and f_r^{spc}. Given the random variables $\mathbf{u}$ and $\mathbf{v}$, the mutual information is Kullback–Leibler (KL) divergence between their joint and marginal distributions as $I(\mathbf{u},\mathbf{v}) = KL(P(\mathbf{u},\mathbf{v}); P(\mathbf{u})P(\mathbf{v}))$. The upper bound of mutual information $I(\mathbf{u},\mathbf{v})$ can computed using vCLUB [52]. At last, we can decrease the correlation among different features by minimizing the unbiased estimation $\hat{I}(\mathbf{u},\mathbf{v})$ [52]. More details are referred to [53]. This allows for more stable training than generative adversarial networks.

In this paper, $\mathbf{u},\mathbf{v} \in \{f_l^{spc}, f_r^{spc}, f_{lr}^{inv}\}$. What this intuitively feels like is making sure that these three features are as far away from each other as possible. Inspired by [52], we apply the $\mathcal{L}_{MIM}$ to all three representations $f_l^{spc}, f_r^{spc}, f_{lr}^{inv}$:

$$\mathcal{L}_{MIM} = \hat{I}(f_{lr}^{inv}, f_l^{spc}) + \hat{I}(f_{lr}^{inv}, f_r^{spc}) \tag{7}$$

where $\hat{I}$ represents the unbiased estimation for vCLUB as described in [52]. Note that f_l^{spc} and f_l^{spc} have already been subjected to distance constraints through the $\mathcal{L}_{Ort}$ in Section 3.2, there is no need to apply additional constraints to these two features in the $\mathcal{L}_{MIM}$.

3.4. Channel-differential representation learning

In addition to channel-specific and invariant features, another important feature is the differential information between two channels, that realized by the channel-differential encoder in Fig. 3(a). It is worth noting that channel-differential encoding is conducted directly on the spectral parameters rather than relying on high-level features. This approach allows for the exploration of differences between the two channels within the original data. Unlike high-level features that tend to offer a condensed representation of the original information, this method preserves the intricate details that might be otherwise obscured.

Specifically, we first calculate the difference $X_{lr}^{\Delta spec}$ between the two spectral parameters X_l^{spec} and X_r^{spec}, and then use the self-attention based network to encode the difference information at high-level, and the resulting output feature is the channel-differential feature f_{lr}^{dif} we need.

$$\begin{aligned} f_{lr}^{dif} &= \text{Self-Attention}(X_{lr}^{\Delta spec}) \\ X_{lr}^{\Delta spec} &= |X_l^{spec} - X_r^{spec}| \end{aligned} \tag{8}$$

It is important to note that temporal information is crucial for the ADD task [1]. Our model mainly considers temporal information in the learning of channel-differential representation, while the channel-specific and -invariant representation learning modules mainly focus on information decoupling. For channel-differential representation learning, we first calculate the difference between the left and right spectral parameters and then utilize the global contextual learning ability of the self-attention mechanism [43] to extract bidirectional temporal information from the differentiating features. This allows us to better explore the temporal information in the speech signal and perform accurate audio deepfake detection.

3.5. Multi-space feature aggregation

The multi-space feature aggregation module aims to fuse information from various spaces to form a joint feature f_{joi}. The feature representations in different spaces do not contribute to a given audio detection to the same extent, therefore, a dynamic weight fusion strategy is established to fuse the four spatial features $f_l^{spc}, f_r^{spc}, f_{lr}^{inv}$, and f_{lr}^{dif}. Specifically, we employ a set of trainable dynamic weight parameters $\mathcal{W}_l^{spc}$ and $\mathcal{W}_l^{spc}$ for the channel-specific representation, $\mathcal{W}_{lr}^{inv}$ for the channel-invariant representation, and $\mathcal{W}_{lr}^{dif}$ for the channel-differential representation respectively.

$$\begin{aligned} f_{joi} =& \mathcal{W}_l^{spc} \cdot f_l^{spc} + \mathcal{W}_r^{spc} \cdot f_r^{spc} + \\ & \mathcal{W}_{lr}^{inv} \cdot f_{lr}^{inv} + \mathcal{W}_{lr}^{dif} \cdot f_{lr}^{dif} \end{aligned} \tag{9}$$

3.6. ADD classifier

At last, the joint feature f_{joi} will be fed into the ADD classifier for bonafide or fake speech prediction. It is important to note that our classifier can be designed as either a frame-level CLS_{fra} or an utterance-level classifier CLS_{utt}. Frame-level classifier CLS_{fra} is suitable for partial ADD [54], while utterance-level classifier CLS_{utt} is suitable for a variety of general-purpose tasks such as speaker verification, speech replay detection, and synthetic speech detection. In particular, frame-level classifier consists of 1D-CNN, FC layer, and sigmoid activation function to calculate Binary Cross-Entropy (BCE) loss $\mathcal{L}_{BCE}$, while utterance-level classifier involves the FC layer to calculate the Weighted Cross-Entropy (WCE) loss $\mathcal{L}_{WCE}$. These two functions are collectively known as the classification loss $\mathcal{L}_{CLS}$. In other words, by switching the type of ADD classifier, we can realize either frame-level or utterance-level ADD tasks.

3.7. Optimization

As mentioned before, MAE, CCE, and orthogonal Losses, that are $\mathcal{L}_{MAE}$, $\mathcal{L}_{CCE}$ and $\mathcal{L}_{Ort}$, act on the channel-specific encoder. MIM loss $\mathcal{L}_{MIM}$ acts on the channel-invariant encoder. classification loss $\mathcal{L}_{CLS}$ acts on the ADD classifier. Therefore, the final training objective of MSCR-ADD can be written as:

$$\begin{aligned} \mathcal{L} =& \lambda_{mae} \cdot \mathcal{L}_{MAE} + \lambda_{cce} \cdot \mathcal{L}_{CCE} + \\ & \lambda_{op} \cdot \mathcal{L}_{Ort} + \lambda_{mim} \cdot \mathcal{L}_{MIM} + \mathcal{L}_{CLS} \end{aligned} \tag{10}$$

where λ_{mae}, λ_{cce}, λ_{op} and λ_{mim} are weighting factors to balance different loss terms. The entire system is trained in an end-to-end manner with well-tuned weighting parameters.

Through the learning of multi-space channel representations, our MSCR-ADD model can learn more comprehensive and robust feature representations for the dual-channel signals during mono-to-binaural conversion, which can provide a more detailed description of the authenticity cues of the input mono speech, and ultimately achieve accurate ADD performance.

4. Experimental setup

In this section, we first describe five benchmark ADD datasets and the M2B dataset in our experiments. Following that, we illustrate implementation details, various advanced baselines, and evaluation metrics.

4.1. Datasets

Our experiments for utterance-level ADD include three datasets: ASVspoof2019 [55], ASVspoof2021 [56] and VSDC [57]. PartialSpoof [54] is used for frame-level ADD experiments.

ASVspoof2019 [55] dataset is released by the ASVspoof2019 challenge, which includes a large-scale dataset containing two subsets of logical access (LA) and physical access (PA). For the LA subtask, 2,580 genuine and 22,800 spoofed speech utterances generated by one of 6 TTS/VC algorithms are used for training. The same spoofing algorithms in the training set are used to create the development set, while the algorithms to generate the evaluation dataset are different. For the PA task, the training set contains 5,400 genuine speech and 48,600 replay spoofing speech comprising 9 different replay configurations (3 categories of attacker-to-speaker recording distance times 3 categories of loudspeaker quality). The evaluation set for the PA task has the same replay spoofing manner as training and development data, with different acoustic configurations. More details are referred to[1].

ASVspoof2021 [56] dataset is released by the ASVspoof2021 challenge, which includes LA, PA, and Deep Fake (DF) tracks. The training and development partitions are the same as ASVspoof2019. Note that there is a new set of evaluation data for each task. For the DF task, the data was primarily generated by ASVspoof2019 LA eval data going through several unknown compression codec channels, which resulted in a total of 611,829 trials. As for the LA task, the TTS, VC and hybrid spoofing attacks were the same as those from the ASVspoof 2019 LA evaluation partition. More details are referred to [58].

VSDC [57] is designed for single- and multi-order voice replay and cloned replay attack detection for diverse and challenging scenarios. VSDC comprises both the first- and second-order replay audios against the bonafide ones. The dataset partition also respects [57,59].

PartialSpoof[2] [54] is constructed based on the ASVspoof2019 LA database by randomly inserting spoof audio segments in pristine audio waveforms. The training, validation, and testing sets are distributed according to the original dataset allocation, consisting of 25,380, 24,844, and 71,237 utterances respectively.

In addition, a <mono, binaural> paired audio data with appropriate size is a prerequisite for M2B converter pretraining. We follow [18] and conduct pretraining on a total of 2 h of paired mono and binaural data[3]. More detailed descriptions are referred to [27].

4.2. Implementation details

For M2S converter pretraining, we train it for 100 epochs using an Adam optimizer. It is worth mentioning that the original M2S converter takes a mono audio segment with a length of 9600 samples as input and then outputs the left and right channels with the same length. To achieve utterance-level M2B conversion, we split each audio into various segments with 64600 samples as the input, then merge the output to obtain the utterance-level left and right channels. During conversion, the conditioning temporal signals C for each audio are randomly sampled from the dataset to avoid the effect of this parameter on the ADD results [27]. The LFCC features follow the official baseline provided in the ASVspoof 2019 [60], extracted with 20 ms window length, 512 FFT points, and 20 filters with their delta and double delta coefficients, making 80-dimensional feature vectors. In the channel-specific encoder, we use *tiny.en* variant of the Whisper model[4]. The N_{CSE} and N_{CIE} are set to 3. The embedding dimension/feed-forward dimension/attention heads in each Transformer layer are set to 1024/4096/16 respectively. Note that we use a dropout of $p = 0.1$ after the self-attention block within our model. The weighting factors in Eq. (10) are set to 0.01/0.005/0.005/0.005 respectively. The entire system is trained for 400 epochs using Adam optimizer [61], where the learning rate is warmed up to a peak of 0.001 for the first 100 updates and then linearly decayed. No data augmentation strategy is used during training. All the hyper-parameters in our systems are tuned on the validation set. All results reported in this paper are average results in addition to the best result from three runs with different random seeds.

4.3. Baselines

To evaluate the performance of our proposed MSCR-ADD, we implement the following state-of-the-art ADD systems as the baselines. For utterance-level ADD, we develop 9 baselines including ResNet [62], LCNN [63], Siamese CNN [64], SE-Res2Net50 [65], Capsule network [66], MCG-Res2Net50+CE [67], Res-TSSDNet [68], RawGAT-ST [69] and M2S-ADD [18].

ResNet [62] based architecture used in ASVspoof 2019 achieved great performance in both PA and LA subtasks [70,71]. We follow [45] and adopt the ResNet18 model consisting of 8 residual blocks. The average pooling and three fully connected (FC) layers are combined as the classifier.

LCNN [63] is the best system in ASVspoof 2017 [72]. The specific characteristic of Light CNN architecture [72] is the usage of the Max-Feature-Map activation (MFM) which is used after each convolution (Conv) operation to choose features that are essential for task solving. We follow [62] and employ a 29-layer LCNN structure.

Siamese CNN [64] contains two identical 1-D CNNs. The input of the convolutional layer is log-probabilities calculated separately by two GMMs trained on genuine and spoofed speech respectively. The branches of CNNs are trained simultaneously and they output two same dimension embedding vectors. Then we concatenate these two vectors and input it to the fully connected layer with dropout and softmax output.

SE-Res2Net50 [65] is a variant of ResNet, that aims at improving multi-scale representation by increasing the number of available receptive fields. This is achieved by connecting smaller filter groups within one block in a hierarchical residual-like style. In addition, the squeeze-and-excitation (SE) block [73] adaptively re-calibrates channel-wise feature responses by explicitly modeling the interdependencies between channels.

Capsule network [66] with a new dynamic routing algorithm is proposed to pay more attention to the local location, thus it is helpful for the audio spoofing detection task. The philosophy of the capsule network is to use a set of capsules to represent the distinct properties of an entity, while the part-whole relationship is implicitly learned through the routing algorithm [74].

MCG-Res2Net50 [67] is an extension of Res2Net. It proposed a multi-group channel-wise gate (MCG) mechanism, which modifies Res2Net to enable a channel-wise gating mechanism in the connection between feature groups. This gating mechanism dynamically selects channel-wise features based on the input, to suppress the less relevant channels and enhance the detection generalizability.

Res-TSSDNet [68] is a light-weight and powerful raw waveform based end-to-end synthetic speech detection network. It considers two types of advanced CNN structures, including ResNet style skip connection [45] and Inception-style parallel convolutions [75], respectively.

RawGAT-ST [69] is also a raw waveform based end-to-end ADD model. Its principal contribution is a spectro-temporal graph attention network (GAT) which learns the relationship between cues spanning different sub-bands and temporal intervals. Using a model-level graph fusion of spectral (S) and temporal (T) sub-graphs and a graph pooling strategy to improve discrimination.

[1] https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf

[2] https://datashare.ed.ac.uk/handle/10283/3336

[3] https://github.com/facebookresearch/BinauralSpeechSynthesis/releases/download/v1.0/binaural_dataset.zip

[4] https://huggingface.co/openai/whisper-tiny.en

M2S-ADD [18] is a novel ADD model, that attempts to discover audio authenticity cues during the mono-to-stereo conversion process. It first projects the mono raw waveform to a binaural signal using a pre-trained binaural audio synthesizer, then employs a dual-branch neural architecture to process the left and right channel signals, respectively. In this way, it effectively reveals the artifacts in the fake audio, thus improving the ADD performance.

For frame-level ADD, we prepare 3 baselines that are LCNN-BLSTM (a) [54], LCNN-BLSTM(b) [54], SELCNN-BLSTM [76].

LCNN-BLSTM(a) [54]: adopt LCNN as the backbone, followed by BiLSTM layers to learn the global temporal sequence context knowledge, since the convolution in an LCNN has a fixed receptive field.

SELCNN-BLSTM [76]: is also treat LCNN as the backbone. It inserts SE blocks [73] into LCNN to enhance the capacity of hidden feature selection. Then, the BiLSTM layer is as the classifier to make the final decision.

LCNN-BLSTM(b) [54]: is another LCNN-based model for partially-spoofed scenario. Unlike LCNN-BLSTM(a), LCNN-BLSTM(b) utilized wav2vec 2.0 [77] based feature to enhance frame-level detection capability.

The type of input signals and front-end features are listed in Table 1. More details about the experimental configurations of baselines are referred to their corresponding reference literature. For a fair comparison, all baselines are trained with one A100 GPU card.

4.4. Evaluation metrics

For utterance-level ADD, we use two metrics: the default minimum tandem detection cost function (min t-DCF) [78,79] and the equal error rate (EER) [80]. The min t-DCF shows the impact of spoofing and the spoofing detection system upon the performance of an automatic speaker verification system, whereas the EER reflects purely standalone spoofing detection performance. We follow [56] and only consider EER for performance evaluation for the DF track of ASVspoof2021. For frame-level ADD, we add other three metrics including precision, recall, and F1 score [81]. All metrics are computed based on frame-level authenticity labels of the partially spoofed audio.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{F1} &= \frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision}+\text{Recall}} \end{aligned} \tag{11}$$

where TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative samples, respectively.

5. Results and discussion

In this work, we propose a novel ADD framework, termed MSCR-ADD, for multi-space channel representation learning of a Mono-to-Binaural conversion based ADD scheme. To validate the MSCR-ADD, we first conduct comparative experiments with currently advanced approaches, investigating the overall performance in terms of utterance or frame-level ADD. Then we conduct the coarse-grained ablation study to verify the effectiveness of the multi-space channel representation learning strategy. Next, we design the fine-grained ablation study to reveal the necessity of some new components in MSCR-ADD, including *Whisper Encoder, Multi-level Channel Classifier, Orthogonal Loss, Mutual Information Maximization Loss and Multi-space Feature Aggregator*, etc. After that, we further compare our proposed method with the related works with similar sprites. We also conduct further analysis of the implications of M2B conversion and the generalization of MSCR-ADD. To qualitatively analyze the improvement of our proposed method, we further visualize the latent representations.

5.1. Main results

We report the main results in terms of utterance and frame-level ADD by switching the types of ADD classifier as mentioned in Section 3.6.

Utterance-level ADD : Table 1 presents the utterance-level detection performance of different approaches. From these experiments results, we have the following observations.

First, our MSCR-ADD framework outperforms all baselines and achieves the best performance on all datasets. (1) For ASVspoof2019 PA data, the worst results came from the LCNN system with EER and min-tDCF scores of 4.51 and 0.1053, respectively. The best system of all baselines is M2S-ADD, with scores of 1.03 and 0.0285 for EER and min-tDCF, respectively. Note that our system achieved significant improvement compared to the M2S-ADD, with 1.05 and 0.0149 improvements in EER and min-tDCF, respectively. We also observe a similar phenomenon in the ASVspoof2019 LA data. MSCR-ADD system has the lowest EER and min-tDCF of all systems at 1.08 and 0.0213 respectively; (2) For ASVspoof2021, PA, LA, and DF datasets share similar performance for all metrics. The proposed MSCR-ADD still holds the advantage by a large margin. Specifically, for the PA data, MSCR-ADD obtained 24.39 and 0.6059 EER and DCF scores, which were 2.69 and 0.0667 lower than the best baseline. For the LA data, MSCR-ADD obtained 1.03 and 0.1834 EER and DCF scores, which were 1.18 and 0.0291 lower than the best baseline. For the DF data, MSCR-ADD obtained the lowest EER of 12.85. (3) For VSDC, our method also performs great improvement in terms of EER and min-tDCF. All the above observations demonstrate that our method exposes more cues of fake audio after converting single-channel mono audio to dual-channel binaural speech. Then, learning a comprehensive multi-space channel feature representation fully utilizes the exposed cues to accurately identify the fake mono audio, which is not considered by other baselines.

Second, our approach mines more comprehensive acoustic information from dual-channel binaural signals. (1) Compared with mono based baselines, the binaural based systems, that are M2S-ADD and our MSCR-ADD, achieved better performance. The last two rows of Table 1 show that the system with binaural as input achieved better EER and DCF scores compared to all other baselines. These results verify that the binaural based systems can handle dual-channel binaural signals and adopt multi-branch architecture to learn the subtle information among the left and right channel signals, thus effectively discovering audio authenticity cues during the mono-to-binaural conversion process. (2) Compared with binaural based baseline, our MSCR-ADD has achieved a further improvement over M2S-ADD. We argue that the M2S-ADD does not explicitly model the channel-specific, invariant, and differential knowledge, which are important for binaural signal understanding. Note that our MSCR-ADD learns such multi-space channel representation, resulting in better ADD performance. (3) Compared with raw audio based baselines, our method uses spectrum as input features, and the details of the spectrum are more favorable for us to discover some details of the fake audio compared to the audio waveform. Although the systems using raw audio as input in baselines perform some advantages over the systems using spectrum input, our method designs a powerful multi-space channel representation learning method for spectrum, which can fully utilize the unique advantages of spectrum feature.

Third, our MSCR-ADD has better robustness in facing complex acoustic environments. As mentioned in Section 4.1, the training and validation sets of ASVspoof2021 are the same as those of 2019, but the test sets for the LA and DF tasks are different. Therefore, the results of LA and DF tasks of ASVspoof2021 can be considered as cross-dataset evaluation. In addition, the test set in 2021 has a more diverse speech environment, such as going through several unknown compression codec channels [56]. Therefore, the results of LA and DF can be used to validate the robustness of our method. From the results in columns

Table 1
Comparison of ADD performance for MSCR-ADD and all advanced baselines in terms of utterance-level. The best performance is highlighted in bold.

System	Input signal	Front-end	ASVspoof2019				ASVspoof2021						VSDC	
			PA		LA		PA		LA		DF			
			EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
ResNet	Mono	Spectrum	3.83	0.0993	3.72	0.1190	40.27	0.9536	10.23	0.6932	23.58	NA	0.98	0.0521
LCNN	Mono	Spectrum	4.51	0.1053	5.06	0.1000	39.54	0.9321	9.68	0.6811	21.36	NA	0.87	0.0503
Siamese CNN	Mono	Spectrum	3.27	0.0987	3.79	0.0930	37.32	0.8857	8.59	0.6525	20.01	NA	0.88	0.0482
SE-Res2Net50	Mono	Spectrum	3.21	0.0825	2.50	0.0743	35.01	0.8615	8.01	0.5415	18.48	NA	0.76	0.0453
Capsule network	Mono	Spectrum	2.95	0.0756	1.97	0.0538	34.86	0.8423	7.35	0.5028	17.26	NA	0.69	0.0388
MCG-Res2Net50+CE	Mono	Spectrum	2.01	0.0752	1.78	0.0523	33.37	0.7889	6.02	0.4513	16.87	NA	0.65	0.0362
Res-TSSDNet	Mono	Raw Audio	1.80	0.0673	1.64	0.0481	29.21	0.6958	5.88	0.4006	16.21	NA	0.60	0.0319
RawGAT-ST	Mono	Raw Audio	1.35	0.0426	1.39	0.0386	28.15	0.6835	3.76	0.2858	15.86	NA	0.55	0.0277
M2S-ADD	Binaural	Raw Audio	1.03	0.0285	1.32	0.0357	27.08	0.6726	2.21	0.2125	14.24	NA	0.49	0.0251
MSCR-ADD (Ours)	Binaural	Spectrum	**0.48**	**0.0136**	**1.08**	**0.0213**	**24.39**	**0.6059**	**1.03**	**0.1834**	**12.85**	NA	**0.38**	**0.0186**

Table 2
Comparison of ADD performance for MSCR-ADD and all advanced baselines in terms of frame-level. The best performance is highlighted in bold.

System	Input signal	Front-end	PartialSpoof				
			EER	min-tDCF	P	R	F
LCNN-BLSTM(a)	Mono	Spectrum	10.15	0.2158	93.25	78.56	81.64
SELCNN-BLSTM	Mono	Spectrum	8.23	0.1837	90.76	76.23	78.52
LCNN-BLSTM(b)	Mono	Raw audio	4.87	0.1326	94.38	79.59	85.35
MSCR-ADD (Ours)	Binaural	Spectrum	**2.98**	**0.0968**	**96.47**	**82.88**	**88.13**

10 to 12 of Table 1, it can be seen that our method performs better than all other baseline systems on the 2021 data, with an EER of 1.03 on LA, a min-tDCF of 0.1834 on LA, and an EER of 12.85 on DF. This indicates that our method exhibits better robustness. Our method can more comprehensively extract acoustic characteristics that are more relevant to speech forgery from speech signals.

Frame-level ADD: Table 2 presents the frame-level detection performance of different approaches. From these experiments results, we have the following observations.

First, our MSCR-ADD framework outperforms all baselines and achieves the best performance on all metrics. Specifically, for EER, the MSCR-ADD achieved 2.98 while all baselines performed higher than 4.0. For min-tDCF, the MSCR-ADD obtained 0.0968, which also gained a remarkable improvement over the baselines. In addition, our MSCR-ADD achieved scores of 96.47, 82.88, and 88.13 for the P, R, and F metrics, respectively. The above results also fully demonstrate that our method uses a multi-space representation of the channel features, which can summarize the information in the spectrum parameters of the two-channel spectrum well after the single-channel mono audio is converted to two-channel binaural audio, and thus can obtain a remarkable performance.

Second, the frame-level multi-space channel representation mechanism is also effective for the frame-level ADD task. The results in Table 2 show that by using a frame-level classifier, our multi-space channel representation can be very effective for the frame-level audio deepfake detection task. The most important thing for frame-level ADD is to discover the details of the fake audio at the frame level, and our model does exactly that.

5.2. Ablation study

We first report the results of a coarse-grained ablation study to analyze the contribution of the novel multi-space channel representation learning strategy. Then, we report the results of the fine-grained ablation study to verify the necessity of various components in MSCR-ADD model. All results contain two parts, which are utterance and frame-level ADD tasks.

5.2.1. Coarse-grained results

Tables 3 and 4 represent the coarse-grained results in terms of utterance and frame-level ADD. There are three parts of ablation that are independent of each other, i.e., each study is conducted where the other two components are kept the same as the MSCR-ADD.

Utterance-level ADD: We first investigate the importance of channel-invariant, specific, and differential representations by discarding each of them. The fifth to seventh rows of Table 3 show the results. We can find that when removing the refined channel-specific representations, the downstream ADD performance degrades a lot under all datasets, which verifies its significance of holding information specific to each channel. For example, the EER of *w/o Channel-Specific Representation* for ASVspoof2019-PA dataset achieves 0.73 and ASVspoof2019-LA achieves 1.26. Other datasets also achieve worse performance. Another metric, or min-tDCF, also performs similar trends on all datasets. For channel-invariant representation, we also observe that the downstream ADD performance drops when discarding it. For example, compared with MSCR-ADD for ASVspoof2019-PA datasets, *w/o Channel-Invariant Representation* drops 0.31 and 0.094 in terms of EER and min-tDCF metrics. Compared with MSCR-ADD for ASVspoof2019-LA datasets, *w/o Channel-Invariant Representation* drops 0.18 and 0.058 in terms of EER and min-tDCF metrics respectively. Similarly, other datasets also achieve worse performance. The observation demonstrates the significance of bridging the channel gap for channel-invariant representations. For channel-differential representation, we also remove it and check the results. In the seventh row of Table 3, we can find that *w/o Channel-Differential Representation* achieves higher results than MSCR-ADD systems, indicating the performance dropping. Take the ASVspoof2019 dataset as an example, the EER and min-tDCF of *w/o Channel-Differential Representation* for ASVspoof2019-PA achieve 0.74 and 0.0228 respectively, while for ASVspoof2019-LA achieve 1.30 and 0.0270 respectively. It proves that the channel-differential representation can be mined for unique feature representations that distinguish them from the other two representations and are used to model subtle differences between channels. In a nutshell, we observe that the new channel-invariant, specific, and differential representations play an important role in MSCR-ADD.

Furthermore, we study the role of channel-invariant, specific, and differential encoders in our MSCR-ADD. We adopt simple feature concatenation and linear projection to replace all three encoders. The last three rows of Table 3 show the results. We observe that all three systems, that are *w/o Channel-Specific Encoder*, *w/o Channel-Invariant Encoder* and *w/o Channel-Differential Encoder*, experience varying degrees of performance degradation. For example, for ASVspoof2019-PA, the EER of min-tDCF of *w/o Channel-Specific Encoder* achieve 0.68 and 0.0198 respectively, *w/o Channel-Invariant Encoder* achieve 0.62 and 0.0209 respectively, while *w/o Channel-Differential Encoder* achieve 0.67 and 0.0195 respectively. In this sense, we conclude that all the encoders in the proposed MSCR-ADD contribute positively to the multi-space channel representation learning.

Table 3
Coarse-grained ablation study on all datasets in terms of utterance-level ADD. Results are reported on six ablation systems. The numbers underlined are copied from Table 1 for easy comparison.

System	ASVspoof2019				ASVspoof2021						VSDC	
	PA		LA		PA		LA		DF			
	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
MSCR-ADD	**0.48**	**0.0136**	**1.08**	**0.0213**	**24.39**	**0.6059**	**1.03**	**0.1834**	**12.85**	NA	**0.38**	**0.0186**
w/o Channel-Specific Representation	0.73	0.0213	1.25	0.0256	26.28	0.6384	1.75	0.2083	13.88	NA	0.45	0.0225
w/o Channel-Invariant Representation	0.79	0.0230	1.26	0.0261	26.57	0.6472	1.80	0.2195	13.91	NA	0.48	0.0232
w/o Channel-Differential Representation	0.74	0.0228	1.30	0.0270	26.73	0.6505	1.83	0.2107	13.99	NA	0.47	0.0235
w/o Channel-Specific Encoder	0.68	0.0198	1.13	0.0225	25.83	0.6132	1.57	0.1945	13.05	NA	0.40	0.0198
w/o Channel-Invariant Encoder	0.62	0.0209	1.15	0.0239	26.07	0.6156	1.59	0.1980	13.16	NA	0.42	0.0199
w/o Channel-Differential Encoder	0.67	0.0195	1.22	0.0231	26.01	0.6202	1.62	0.1976	13.11	MA	0.41	0.0203

Table 4
Coarse-grained ablation study in terms of frame-level ADD. Results are reported on six ablation systems. The numbers underlined are copied from Table 1 for easy comparison.

System	PartialSpoof				
	EER	min-tDCF	P	R	F
MSCR-ADD	**2.98**	**0.0968**	**96.47**	**82.88**	**88.13**
w/o Channel-Specific Representation	3.76	0.1235	96.01	82.35	87.99
w/o Channel-Invariant Representation	3.75	0.1241	95.89	82.20	87.05
w/o Channel-Differential Representation	3.82	0.1243	95.89	82.37	86.98
w/o Channel-Specific Encoder	3.15	0.1023	94.71	80.47	85.87
w/o Channel-Invariant Encoder	3.18	0.1052	94.69	80.89	85.98
w/o Channel-Differential Encoder	3.20	0.1049	94.85	81.05	86.03

Frame-level ADD : Table 4 represents the coarse-grained results in terms of frame-level ADD. Following the previous section, we also investigate the importance of channel-invariant, specific, and differential representations and three encoders. From the results in Table 4, it can be seen that after removing the three representations respectively, the performance of the three systems in all the metrics shows a significant degradation. The performance of the three systems is also worse after replacing the three encoder structures respectively using the simple structure, which indicates that both our three feature representations as well as the encoder structures play an important role in model improvement in frame-level ADD. These three channel representations and encoders extract rich information from binaural fake audio from different perspectives for speech authentication.

5.2.2. Fine-grained results

Tables 5 and 6 represent the fine-grained results in terms of utterance and frame-level ADD. We developed six ablation systems for comparison: (1) ***w/o Whisper Encoder***: we remove the whisper encoder and the channel-specific encoder just takes the spectrum feature as input; (2) ***w/o Whisper Encoder Fine-tuning***: we adopt the pre-trained whisper encoder for feature extraction while not update it during MSCR-ADD training; (3) ***w/o Multi-level Channel Classifier***: we discard the multi-level channel classifier. It means the channel-specific features of the two channels lose the constraints in terms of distance; (4) ***w/o Orthogonal Loss***: Unlike previous system, we just remove the $\mathcal{L}_{Ort}$; (5) ***w/o Mutual Information Maximization Loss***: we remove the $\mathcal{L}_{MIM}$; (6) ***w/o Multi-space Feature Aggregator***: we just employ feature concatenation on various channel representations to obtain the joint feature. There are six parts of ablation that are independent of each other, i.e., each study is conducted where the other five components are kept the same as the MSCR-ADD.

Utterance-level ADD: Table 5 shows that all modules play a vital role in the performance of the MSCR-ADD model in terms of utterance-level ADD. For example, when removing the whisper encoder, the EER of ASVspoof2019-PA is 0.53, which is higher than MSCR-ADD by a large margin. We can find that the whisper encoder allows us to extract the high-level channel feature that is helpful for the ADD task. After stopping to fine-tune the whisper encoder, the EER of ASVspoof2019-PA is 0.52, which also performs performance drops compared with MSCR-ADD. It further shows that fine-tuning the whisper encoder designed for the speech recognition task for our ADD task is necessary to make our extracted features more suitable for the ADD task. When removing the multi-level channel classifier in the channel-specific encoder, the EER of ASVspoof2019-PA obtains 0.52. It is also well demonstrated that our method can effectively enhance the frame-level discriminability of learned channel-specific representation, thus helping to obtain meaningful channel-specific representation. Similarly, after deleting the $\mathcal{L}_{Ort}$, the EER of ASVspoof2019-PA obtains 0.54, which achieves further performance degradation. We believe that orthogonal Loss can constrain the channel-specific representation of two channels well enough to obtain two robust channel-specific representations. Mutual information maximization loss is used to obtain robust channel-invariant features, and after removing it, the performance of the model shows a degradation. Therefore, it is also reasonable to believe that $\mathcal{L}_{MIM}$ is indispensable for the learning of channel-invariant representations. At last, by discarding the multi-space feature aggregator, the EER of ASVspoof2019-PA obtains 0.54. This indicates that weight-based fusion of different channel features can effectively fuse the contributions of different information to obtain meaningful joint features, while simple feature concatenation cannot do so. To summarize, all our modules are designed for MSCR-ADD, which is crucial for speech authentication. The same conclusion can be drawn from the performance of other datasets for all metrics.

Frame-level ADD: Table 6 also demonstrates that all modules are integral to the performance of the MSCR-ADD model in terms of frame-level ADD. Similar to the results of Table 5, we can see that the performance of the model decreases to varying degrees after the removal of each of the six modules. This shows the effectiveness of the methods we designed in different aspects. Overall, these methods play an important role in learning a comprehensive representation of features in binaural information for M2B-based ADD tasks.

5.3. Comparison with related works

As illustrated in Section 2.3, prior works have investigated the multi-space feature learning. We have prepared a typical representative, that is MISA [33], of the relevant field for comparison. For a fair comparison, we implement it in our framework as a comparison, where we employ the *Similarity Loss* and *Difference Loss* to constraint the channel-specific and invariant representation learning while keeping channel-differential representation the same as MSCR-ADD.

Tables 7 and 8 report the results in terms of utterance and frame-level ADD. By observing the results of the experiment, we conclude two observations. (1) Multi-space channel representation learning does learn more comprehensive feature representations, which are crucial for mono-to-binaural based ADD. We compare the results in Tables 1 and 2 with those in Tables 7 and 8, and we can see that both MISA and our model are significantly improved compared to other ADD baselines, significantly improving the accuracy of ADD. This demonstrates the

Table 5
Fine-grained ablation study on all datasets in terms of utterance-level ADD. Results are reported on six ablation systems. The numbers underlined are copied from Table 1 for easy comparison.

System	ASVspoof2019				ASVspoof2021						VSDC	
	PA		LA		PA		LA		DF			
	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
MSCR-ADD	**0.48**	**0.0136**	**1.08**	**0.0213**	**24.39**	**0.6059**	**1.03**	**0.1834**	**12.85**	NA	**0.38**	**0.0186**
w/o Whisper Encoder	0.53	0.0185	1.22	0.0305	26.27	0.6632	1.98	0.2059	13.54	NA	0.40	0.0225
w/o Whisper Encoder Fine-tuning	0.52	0.0191	1.21	0.0298	26.31	0.6529	1.95	0.2086	13.57	NA	0.41	0.0226
w/o Multi-level Channel Classifier	0.52	0.0187	1.23	0.0301	26.29	0.6530	2.03	0.1987	13.49	NA	0.39	0.0231
w/o Orthogonal Loss	0.54	0.0193	1.20	0.0299	25.88	0.6612	2.09	0.2015	14.01	NA	0.42	0.0219
w/o Mutual Information Maximization Loss	0.55	0.0195	1.30	0.0312	25.93	0.6607	1.99	0.2024	13.72	NA	0.44	0.0220
w/o Multi-space Feature Aggregator	0.54	0.0198	1.29	0.0306	26.05	0.6593	2.08	0.2007	13.68	NA	0.41	0.0229

Table 6
Fine-grained ablation study on all datasets in terms of utterance-level ADD. Results are reported on six ablation systems. The numbers underlined are copied from Table 1 for easy comparison.

System	PartialSpoof				
	EER	min-tDCF	P	R	F
MSCR-ADD	**2.98**	**0.0968**	**96.47**	**82.88**	**88.13**
w/o Whisper Encoder	3.88	0.1257	94.85	80.11	86.05
w/o Whisper Encoder Fine-tuning	3.90	0.1249	94.97	80.23	86.12
w/o Multi-level Channel Classifier	3.89	0.1198	94.52	80.02	86.01
w/o Orthogonal Loss	3.94	0.1205	94.38	79.94	85.87
w/o Mutual Information Maximization Loss	3.96	0.1223	94.50	80.01	85.99
w/o Multi-space Feature Aggregator	3.75	0.1218	94.68	80.09	86.10

effectiveness of the multi-space channel representation learning strategy; (2) Our proposed method is more robust for multi-space channel representation learning than MISA. As can be seen from Tables 7 and 8, the results of the MISA model also show a significant decrease on different datasets, which further fully demonstrates the effectiveness of our method in multi-space channel representation learning.

5.4. Discussion of the M2B conversion

Note that the first step of the pipeline of our MSCR-ADD is to convert single-channel mono signal to binaural. Therefore, the impact of M2B is discussed. We prepared another ADD model, that is *MSCR-ADD with M2B.v2* for comparison. For *MSCR-ADD with M2B.v2*, we replace the M2B converter in MSCR-ADD with another powerful binaural audio synthesizer [19], where a two-stage conditional diffusion probabilistic model is proposed to achieve remarkable mono to binaural conversion performance.

The last rows of Tables 7 and 8 report the comparison results regarding utterance and frame-level ADD. We can see that despite replacing the M2B converter, the performance on the different data is still remarkable. For example, the utterance-level EER and min-tDCF of ASVspoof2019-PA for *MSCR-ADD with M2B.v2* achieve 0.49 and 0.0139. Frame-level EER and min-tDCF achieve 3.01 and 0.981 respectively. There is still a clear advantage over the baselines in Tables 1 and 2. The above results fully demonstrate that our method can be very effective in mining cues of fake audio in dual-channel binaural signals for accurate ADD without being affected by the M2B converter.

5.5. Discussion of the layer number of channel-specific/invariant encoders

In this section, we investigate the impact of hyperparameter settings on the experimental results in the model. We choose the number of layers in the channel-specific and invariant encoders as the focus of our hyperparameter experiment. We set this number to 1, 2, 3, 4, and 5, respectively, to observe the final results. Tables 9 and 10 display the results at sentence-level and frame-level, respectively.

From Table 9, it can be seen that our model shows a trend of improvement followed by deterioration as the number of layers increases. The best performance is achieved when the number of layers equals 3. For example, in the PA task of ASVspoof2019, when the number of layers is 1, the EER is 0.61, when it is 2, the EER is 0.52, and when it is 3, the EER reaches 0.48, achieving the best result. However, as the number of layers increases to 4 and 5, the EER rises again, indicating a gradual decline in performance. Similar trends can be observed for other metrics in other datasets. Similarly, from the results in Table 10, it can be observed that the best performance is achieved when the number of layers equals 3. These results suggest that as the number of encoder layers increases, the model can learn the channel-specific feature representation well. If there are too few layers, the learning ability of the encoder is insufficient, and if there are too many layers, the learned features may ignore the original acoustic information of the speech. Therefore, the best performance is achieved when the number of encoder layers equals 3.

5.6. Visualization study

This section seeks to conduct a visualization study to validate the MSCR-ADD. We prepare four systems, that are MISA, *w/o Orthogonal Loss*, *w/o Mutual Information Maximization Loss* and *MSCR-ADD*, for comparison since these two modules play an important role in constraints on multi-space channel representation learning.

Fig. 4 presents the t-SNE visualization of channel-specific, invariant and differential representations of various systems. The blue and green points denote channel-specific representations for left and right channels, respectively, and the orange points denote channel-invariant representations. The red points denote channel-differential representations. By checking the visualization results, we have the following observations.

First, the feature distributions in different spaces obtained by our MSCR-ADD method form a very clear clustering effect, which is significantly better than MISA baseline. For example, the four feature points in subfigure (d) form clear clusters, and none of the feature representations in different spaces overlap with those in other spaces. However, the four feature points in subfigure (a) do not form clear boundaries, and some of the points overlap to form fuzzy boundaries in terms of the multi-space concept. The reason behind the above results still lies in the fact that our method learns to explicitly define and constrain the features in different spaces, learns the robust feature representations for respective spaces, and ultimately develops clear multi-space channel representation.

Second, the clustering effect of the features becomes worse when the two key losses are removed, indicating that these two losses are crucial for multi-space channel representation learning, which further validates the previous conclusion. For example, comparing subfigure (d) with subfigures (b) and (c), the clustering of feature points in subfigures (b) and (c) is not as pronounced as in subfigure (d), and the boundaries of some categories are blurred. But compared with (a), it still shows some clustering effect. The above results once again emphasize that our channel-specific and invariant encoders involve the above two losses

Table 7
Comparison between MSCR-ADD and related works, and discussion about the implications of M2B conversion about the multi-space feature learning on all datasets in terms of utterance-level ADD. The numbers underlined are copied from Table 1 for easy comparison.

System	ASVspoof2019				ASVspoof2021						VSDC	
	PA		LA		PA		LA		DF			
	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
MSCR-ADD	**0.48**	**0.0136**	**1.08**	**0.0213**	**24.39**	**0.6059**	**1.03**	**0.1834**	**12.85**	NA	**0.38**	**0.0186**
Comparison with related works												
MISA	0.51	0.0214	1.29	0.0352	27.22	0.6734	2.01	0.2092	14.02	NA	0.45	0.249
Discussion of the M2B Conversion												
MSCR-ADD with M2B.v2	0.49	0.0139	1.12	0.0221	25.13	0.6111	1.08	0.1921	12.99	NA	0.40	0.188

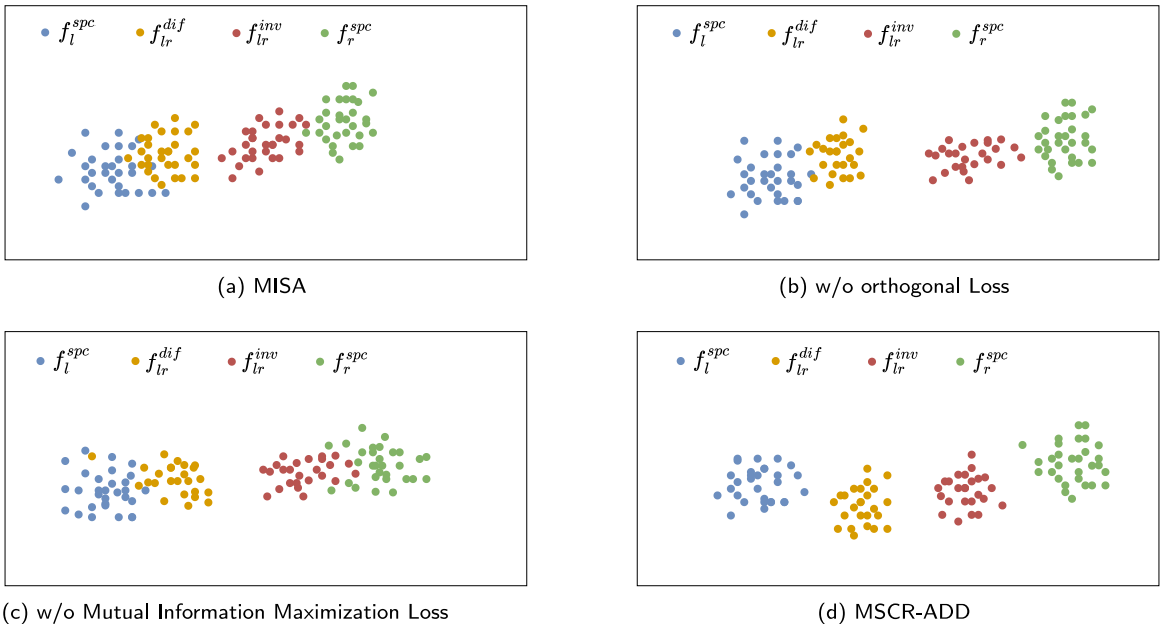

Fig. 4. The t-SNE visualization of channel-invariant, -specific, and -differential representations from (a) MISA, (b) w/o Orthogonal Loss, (c) w/o Mutual Information Maximization Loss, and (d) our MSCR-ADD. The blue and green points denote channel-specific representations for left and right channels, respectively, and the orange points denote channel-invariant representations. The red points denote channel-differential representations. This study is conducted on frame-level representations using a portion of ASVspoof2019 test set.

Table 8
Comparison between MSCR-ADD and related works, and discussion about the implications of M2B conversion about the multi-space feature learning on all datasets in terms of frame-level ADD. The numbers underlined are copied from Table 1 for easy comparison.

System	PartialSpoof				
	EER	min-tDCF	P	R	F
MSCR-ADD	**2.98**	**0.0968**	**96.47**	**82.88**	**88.13**
Comparison with related works					
MISA	3.99	0.1255	93.01	78.83	84.96
Discussion of the M2B Conversion					
MSCR-ADD with M2B.v2	3.01	0.0981	96.01	82.83	87.02

to constrain the learning of features and learn interpretable and strong multi-space channel representations.

In a nutshell, our model defined three feature spaces and tried to learn the feature representation in all three spaces. We believe that fusing these features will help provide a comprehensive feature representation that is useful for ADD.

6. Conclusions

In this paper, we propose MSCR-ADD, a novel multi-space channel representation learning for mono-to-binaural based ADD, which captures the channel-specific, invariant, and differential features to obtain robust and comprehensive representation for binaural signals. Specifically, we first defined three feature spaces, which are channel-specific, invariant and differential spaces, to provide a holistic view of binaural signals for ADD tasks. Then we proposed channel-specific, invariant and differential encoders to learn the channel-specific, invariant and differential representations for the converted binaural audio. Note that we designed the *Orthogonal Projection* and *Mutual Information Maximization Losses* to further constrain the channel-specific and invariant representations. Finally, channel-specific, invariant, and differential representations are fused to help the final decision. We also achieved utterance and frame-level ADD in our framework. Experiments on four public benchmarks show that our MSCR-ADD outperforms all baselines by a large margin.

Despite the progress made by our method in audio deepfake detection, there are still some limitations that need to be addressed. For example, the performance of our method in detecting forged regions needs further validation. Additionally, it is still unknown whether our method can accurately identify forged singing voices. In the future,

Table 9
Utterance-level ADD results of the comparison experiment on the layer number of Channel-Specific/Invariant Encoders. The numbers underlined are copied from Table 1 for easy comparison.

System	ASVspoof2019				ASVspoof2021						VSDC	
	PA		LA		PA		LA		DF			
	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
N_{CSE}, N_{CIE} = 1	0.61	0.0159	1.31	0.0293	27.03	0.6802	1.57	0.2135	13.81	NA	0.55	0.0302
N_{CSE}, N_{CIE} = 2	0.52	0.0146	1.20	0.0278	26.55	0.6675	1.50	0.2102	13.09	NA	0.50	0.0285
N_{CSE}, N_{CIE} = 3	**0.48**	**0.0136**	**1.08**	**0.0213**	**24.39**	**0.6059**	**1.03**	**0.1834**	**12.85**	NA	**0.38**	**0.0186**
N_{CSE}, N_{CIE} = 4	0.52	0.0148	1.22	0.0245	25.32	0.6388	1.39	0.2033	13.08	NA	0.49	0.0233
N_{CSE}, N_{CIE} = 5	0.58	0.0157	1.28	0.0262	25.84	0.6404	1.36	0.2018	13.45	NA	0.42	0.0230

Table 10
Frame-level ADD results of the comparison experiment on the layer number of Channel-Specific/Invariant Encoders. The numbers underlined are copied from Table 1 for easy comparison.

System	PartialSpoof				
	EER	min-tDCF	P	R	F
N_{CSE}, N_{CIE} = 1	3.78	0.1158	94.21	79.05	86.12
N_{CSE}, N_{CIE} = 2	3.56	0.1134	94.59	80.53	86.88
N_{CSE}, N_{CIE} = 3	**2.98**	**0.0968**	**96.47**	**82.88**	**88.13**
N_{CSE}, N_{CIE} = 4	3.29	0.1127	95.04	81.13	86.87
N_{CSE}, N_{CIE} = 5	3.21	0.1125	95.56	81.52	86.99

we will explore the applicability of our method in more diverse scenarios. Furthermore, we will continue to explore more comprehensive channel feature representations and further disentangling of feature representations from various spaces.

CRediT authorship contribution statement

Rui Liu: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Validation, Writing – original draft, Writing – review & editing. **Jinhua Zhang:** Investigation, Resources. **Guanglai Gao:** Methodology,Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rui Liu reports financial support was provided by Inner Mongolia University. Rui Liu reports a relationship with Inner Mongolia University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62206136), Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B1212010004),the High-level Talents Introduction Project of Inner Mongolia University (No. 10000-22311201), and the "Inner Mongolia Science and Technology Achievement Transfer and Transformation Demonstration Zone, University Collaborative Innovation Base, and University Entrepreneurship Training Base" Construction Project (Supercomputing Power Project) (No.21300-231510).

References

[1] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al., Add 2022: the first audio deep synthesis detection challenge, in: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 9216–9220.

[2] Cunhang Fan, Jun Xue, Shunbo Dong, Mingming Ding, Jiangyan Yi, Jinpeng Li, Zhao Lv, Subband fusion of complex spectrogram for fake speech detection, Speech Communication 155 (2023) 102988.

[3] Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, Zhao Lv, Learning from yourself: a self-distillation method for fake speech detection, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, Neural speech synthesis with transformer network, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 6706–6713.

[5] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech: fast, robust and controllable text to speech, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 3171–3180.

[6] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech 2: fast and high-quality end-to-end text to speech, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021, OpenReview.net.

[7] Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al., Mega-TTS: zero-shot text-to-speech at scale with intrinsic inductive bias, 2023, arXiv preprint arXiv:2306.03509.

[8] Rui Liu, Yifan Hu, Haolin Zuo, Zhaojie Luo, Longbiao Wang, Guanglai Gao, Text-to-speech for low-resource agglutinative language with morphology-aware language model pre-training, IEEE/ACM Trans. Audio Speech Lang. Process. (2023) 1–13.

[9] Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, Haizhou Li, Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling, arXiv preprint arXiv:2312.11947 (2023).

[10] Yinghao Aaron Li, Cong Han, Nima Mesgarani, Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models, in: 2022 IEEE Spoken Language Technology Workshop, SLT, IEEE, 2023, pp. 920–927.

[11] Jingyi Li, Weiping Tu, Li Xiao, Freevc: towards high-quality text-free one-shot voice conversion, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[12] Xin Wang, Junichi Yamagishi, A comparative study on recent neural spoofing countermeasures for synthetic speech detection, in: Proc. Interspeech 2021, 2021, pp. 4259–4263.

[13] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, Anthony Larcher, End-to-end anti-spoofing with rawnet2, in: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6369–6373.

[14] Priyanka Gupta, Piyushkumar K. Chodingala, Hemant A. Patil, Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components, Comput. Speech Lang. 77 (2023) 101423.

[15] Bhusan Chettri, Bob L. Sturm, A deeper look at Gaussian mixture model based anti-spoofing systems, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 5159–5163.

[16] Zhuxin Chen, Weibin Zhang, Zhifeng Xie, Xiangmin Xu, Dongpeng Chen, Recurrent neural networks for automatic replay spoofing attack detection, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 2052–2056.

[17] Hemlata Tak, Jee weon Jung, Jose Patino, Massimiliano Todisco, Nicholas Evans, Graph attention networks for anti-spoofing, in: Proc. Interspeech 2021, 2021, pp. 2356–2360.

[18] Rui Liu, Jinhua Zhang, Guanglai Gao, Haizhou Li, Betray oneself: a novel audio DeepFake detection model via mono-to-stereo conversion, in: Proc. INTERSPEECH 2023, 2023, pp. 3999–4003.

[19] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al., Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis, Adv. Neural Inf. Process. Syst. 35 (2022) 23689–23700.

[20] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, Orthogonal projection loss, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12333–12343.

[21] Rui Liu, Berrak Sisman, Haizhou Li, Graphspeech: syntax-aware graph attention network for neural speech synthesis, in: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 6059–6063.

[22] Berrak Sisman, Junichi Yamagishi, Simon King, Haizhou Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2020) 132–157.

[23] Tanvina B. Patel, Hemant A. Patil, Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech, in: Proc. Interspeech 2015, 2015, pp. 2062–2066.

[24] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, Ha-Jin Yu, RawNet: advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification, in: Proc. Interspeech 2019, 2019, pp. 1268–1272.

[25] Zhenchun Lei, Yingen Yang, Changhong Liu, Jihua Ye, Siamese convolutional neural network using gaussian probability feature for spoofing speech detection, in: Proc. Interspeech 2020, 2020, pp. 1116–1120.

[26] Jiakang Li, Meng Sun, Xiongwei Zhang, Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, 2019, pp. 1517–1522.

[27] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, Yaser Sheikh, Neural synthesis of binaural speech from mono audio, in: International Conference on Learning Representations, 2021.

[28] Pablo M. Delgado, Jürgen Herre, Objective assessment of spatial audio quality using directional loudness maps, in: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 621–625.

[29] Tianyun Liu, Diqun Yan, Rangding Wang, Nan Yan, Gang Chen, Identification of fake stereo audio using SVM and CNN, Information 12 (7) (2021).

[30] Bahareh Tolooshams, Kazuhito Koishida, A training framework for stereo-aware speech enhancement using deep neural networks, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021) 6962–6966.

[31] Kranti Kumar Parida, Siddharth Srivastava, Gaurav Sharma, Beyond mono to binaural: generating binaural audio from mono audio with depth and cross modal attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022, pp. 3347–3356.

[32] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, Dahua Lin, Visually informed binaural audio generation without binaural audios, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 15485–15494.

[33] Devamanyu Hazarika, Roger Zimmermann, Soujanya Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[34] Wenmeng Yu, Hua Xu, Ziqi Yuan, Jiele Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 12, 2021, pp. 10790–10797.

[35] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, Haizhou Li, Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[36] Ziyu Wei, Xi Yang, Nannan Wang, Xinbo Gao, Syncretic modality collaborative learning for visible infrared person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 225–234.

[37] Nianchang Huang, Jianan Liu, Yongjiang Luo, Qiang Zhang, Jungong Han, Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification, Pattern Recognit. 135 (2023) 109145.

[38] Xin Hao, Sanyuan Zhao, Mang Ye, Jianbing Shen, Cross-modality person re-identification via modality confusion and center aggregation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16403–16412.

[39] Haixia Xiong, Weihua Ou, Zengxian Yan, Jianping Gou, Quan Zhou, Anzhi Wang, Modality-specific matrix factorization hashing for cross-modal retrieval, J. Ambient Intell. Humaniz. Comput. (2020) 1–15.

[40] Ruoyu Liu, Yao Zhao, Shikui Wei, Liang Zheng, Yi Yang, Modality-invariant image-text embedding for image-sentence matching, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 15 (1) (2019) 1–19.

[41] Zhanxiang Feng, Jianhuang Lai, Xiaohua Xie, Learning modality-specific representations for visible-infrared person re-identification, IEEE Trans. Image Process. 29 (2019) 579–590.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[44] Dan Hendrycks, Kevin Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[46] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 3451–3460.

[47] Steffen Schneider, Alexei Baevski, Ronan Collobert, Michael Auli, wav2vec: unsupervised pre-training for speech recognition, in: Interspeech 2019, ISCA, 2019.

[48] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, Nicholas W.D. Evans, Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, in: The Speaker and Language Recognition Workshop, 28 June - 1 July 2022, Beijing, China, Odyssey 2022, ISCA, 2022, pp. 112–119.

[49] Chenglong Wang, Jiangyan Yi, Jianhua Tao, Haiyang Sun, Xun Chen, Zhengkun Tian, Haoxin Ma, Cunhang Fan, Ruibo Fu, Fully automated end-to-end fake audio detection, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022, pp. 27–33.

[50] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, Dumitru Erhan, Domain separation networks, Adv. Neural Inf. Process. Syst. 29 (2016).

[51] Yuchen Hu, Chen Chen, Ruizhe Li, Heqing Zou, Eng Siong Chng, MIR-GAN: refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition, 2023, arXiv preprint arXiv:2306.10567.

[52] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, Helen Meng, VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion, in: Proc. Interspeech 2021, 2021, pp. 1344–1348.

[53] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, Lawrence Carin, Club: A contrastive log-ratio upper bound of mutual information, in: International Conference on Machine Learning, PMLR, 2020, pp. 1779–1788.

[54] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans, An initial investigation for detecting partially spoofed audio, in: Interspeech 2021, ISCA, 2021, pp. 4264–4268.

[55] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Comput. Speech Lang. 64 (2020) 101114.

[56] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al., ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge, 2021.

[57] Roland Baumann, Khalid Mahmood Malik, Ali Javed, Andersen Ball, Brandon Kujawa, Hafiz Malik, Voice spoofing detection corpus for single and multi-order audio replays, Comput. Speech Lang. 65 (2021) 101132.

[58] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Trans. Audio Speech Lang. Process. (2023).

[59] Ali Javed, Khalid Mahmood Malik, Hafiz Malik, Aun Irtaza, Voice spoofing detector: A unified anti-spoofing framework, Expert Syst. Appl. 198 (2022) 116770.

[60] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Comput. Speech Lang. 64 (2020) 101114.

[61] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[62] Jichen Yang, Hongji Wang, Rohan Kumar Das, Yanmin Qian, Modified magnitude-phase spectrum information for spoofing detection, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 1065–1078.

[63] G. Lavrentyeva, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlov, S. Novoselov, STC antispoofing systems for the AsVspoof2019 challenge, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019, pp. 1033–1037.

[64] Zhenchun Lei, Yingen Yang, Changhong Liu, Jihua Ye, Siamese convolutional neural network using gaussian probability feature for spoofing speech detection., in: Proc. Interspeech 2020, 2020, pp. 1116–1120.

[65] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, Helen Meng, Replay and synthetic speech detection with res2net architecture, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6354–6358.
[66] Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, Z Jane Wang, A capsule network based approach for detection of audio spoofing attacks, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6359–6363.
[67] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, Helen Meng, Channel-wise gated Res2Net: towards robust detection of synthetic speech attacks, in: Proc. Interspeech 2021, 2021, pp. 4314–4318.
[68] Guang Hua, Andrew Beng Jin Teoh, Haijian Zhang, Towards end-to-end synthetic speech detection, IEEE Signal Process. Lett. 28 (2021) 1265–1269.
[69] Hemlata Tak, Jee-Weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, Nicholas Evans, End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection, in: Automatic Speaker Verification and Spoofing Countermeasures Challenge, ASVSPOOF 2021, ISCA, 2021, pp. 1–8.
[70] Yexin Yang, Hongji Wang, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, Kai Yu, The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge, in: Interspeech 2019, ISCA, 2019.
[71] Joao Monteiro, Jahangir Alam, Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU, IEEE, 2019, pp. 1003–1010.
[72] G. Lavrentyeva, S. Novoselov, E. Malykh, O. Kudashev, V. Shchemelinin, A. Kozlov, Audio replay attack detection with deep learning frameworks, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017, pp. 82–86.
[73] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[74] Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 2307–2311.
[75] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2015, pp. 1–9.
[76] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Multi-task learning in utterance-level and segmental-level spoof detection, in: 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, ISCA, 2021.
[77] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst. 33 (2020) 12449–12460.
[78] Tomi Kinnunen, Kong Aik Lee, Hector Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, Douglas A. Reynolds, t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, in: The Speaker and Language Recognition Workshop, Odyssey 2018, ISCA, 2018.
[79] Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, et al., Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 2195–2210.
[80] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, Aleksandr Sizov, ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association, International Speech Communication Association, 2015, pp. 2037–2041.
[81] Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, Ruibo Fu, Half-truth: a partially fake audio detection dataset, in: Proc. Interspeech 2021, 2021, pp. 1654–1658.