IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Dual-Channel Deepfake Audio Detection: Leveraging Direct and Reverberant Waveforms

GUNWOO LEE[1,*], JUNGMIN LEE[1,*], MINKYO JUNG[1], JOSEPH LEE[2], KIHUN HONG[1], SOUHWAN JUNG[1], YOSEOB HAN[1,3]

[1]School of Electronic Engineering, Soongsil University, Seoul, Republic of Korea
[2]Statistics and Actuarial Science, Soongsil University, Seoul, Republic of Korea
[3]Department of Intelligent Semiconductors, Soongsil University, Seoul, Republic of Korea
[*]These authors contributed equally to this work

Corresponding author: Yoseob Han (e-mail: yoseob.han@ssu.ac.kr).

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2020-0-01602) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00263037, Robust deepfake audio detection development against adversarial attacks)

ABSTRACT Deepfake content-including audio, video, images, and text-synthesized or modified using artificial intelligence is designed to convincingly mimic real content. As deepfake generation technology advances, detecting deepfake content presents significant challenges. While recent progress has been made in detection techniques, identifying deepfake audio remains particularly challenging. Previous approaches have attempted to capture deepfake features by combining video and audio content; however, these methods are ineffective when video and audio are mismatched due to occlusion. To address this, we proposes a novel dual-channel deepfake audio detection model that leverages the direct and reverberant components extracted from raw audio signals, focusing exclusively on audio-based detection without reliance on video content. Across various datasets, including ASVspoof2019, FakeAVCeleb, and sport press conference datasets collected by our group, the proposed dual-channel model demonstrates significant improvements in quantitative metrics such as equal error rate and area under the curve. The implementation is available at https://github.com/gunwoo5034/Dual-Channel-Audio-Deepfake-Detection.

INDEX TERMS Deepfake audio detection, dual-channel data, direct waveform, reverberant waveform.

I. INTRODUCTION

In modern society, the advancement of voice technology has become a significant topic, with audio-based applications rapidly expanding due to the progress of deep learning (DL) technology. As speech recognition and synthesis technologies are increasingly integrated into daily life, the emergence of ***Deepfake Voice***—synthetic voices generated by mimicking real human speech using artificial intelligence (AI)—is accelerating. While these technologies offer numerous beneficial applications, they are also vulnerable to misuse for malicious purposes, such as fraud, threats, and the dissemination of false information. For instance, with the growing prevalence of interconnected Internet of Things (IoT) devices, unauthorized access through synthesized voices poses a tangible risk. To address these challenges, researchers have actively developed deepfake audio detection technologies aimed at mitigating the associated risks.

Recent advancements in deep learning (DL) technology have revolutionized multimedia manipulation through the use of generative adversarial networks (GANs), significantly impacting computer vision and deepfake creation. Various face-swapping models, such as FaceSwap [1], FaceShifter [2], Face2Face [3], DeepFaceLab [4], and Neural Textures [5], have been developed to transform faces in original videos into target images. The widespread availability and misuse of these deepfake methods underscore the urgent need for advanced detection techniques to mitigate the challenges posed by their malicious applications.

Recently, numerous convolutional neural network (CNN) architectures have been proposed [6]–[13] to capture spatial image features and/or temporal audio features, such as mel-frequency cepstral coefficients (MFCC) and linear-frequency cepstral coefficients (LFCC), for deepfake detection. Recurrent neural network (RNN) models [14], [15] have also been employed to differentiate between real and fake audio using 1D temporal features. Furthermore, transformer-based

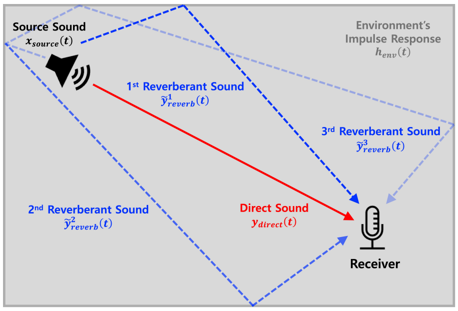

FIGURE 1: Audio physics model containing direct waveform $y_{direct}(t)$ and i_{th} reverberant waveform $\tilde{y}^i_{reverb}(t)$ in a given environment defined by impulse response $h_{env}(t)$. Sum of all reverberant waveforms is defined as $y_{reverb}(t) = \sum \tilde{y}^i_{reverb}(t)$.

architectures [16], [17] incorporating multi-head attention mechanisms have been introduced to enhance detection performance. Other approaches combining CNNs with attention mechanisms have also been developed [11], [18]–[20]. Alternatively, deepfake detection methods that consider both audio and video typically focus on lip-sync analysis [6]–[8]. However, these methods may face challenges in accurately detecting discrepancies between lip movements and sounds when occlusion occurs in the video content.

Deepfake audio is typically generated using text-to-speech (TTS) [21]–[23] and/or voice conversion (VC) [24], [25] techniques. However, these approaches do not account for audio physics models (see Figure 1), such as head movement, recording environment, and audio source, relying solely on the capabilities of the AI model. Consequently, generated deepfake audio struggles to replicate the complexities of audio physics models, particularly the interplay of direct and reverberant sounds that depend on environmental factors.

Similar to deepfake audio generation models [21]–[25], previously proposed deepfake audio detection models [6]–[20] have also been developed without incorporating audio physics models that account for environmental information.

Based on the characteristics of direct and reverberant sounds influenced by environmental factors, we propose a novel dual-channel deepfake audio detection method. As illustrated in Figure 1, direct sound refers to audio arriving directly from the sound source, while reverberant sound is the audio that arrives after being reflected and repeatedly scattered within a space. In real audio, a distinct difference exists between direct and reverberant sounds. However, in deepfake audio, these differences are typically negligible or entirely absent. These differences can be quantified using the direct-to-reverberant ratio (DRR), which will be elaborated on in Section III. Consequently, the proposed method leverages dual-channel audio, focusing on the direct and reverberant sound characteristics, to distinguish between real and deepfake audio based on the principles of the audio physics model. Furthermore, our group has collected a new Sports Press Conference (SPC) dataset, offering longer and higher-quality recordings compared to existing datasets. This dataset facilitates a more comprehensive analysis from various perspectives. The major contributions of this work are as follows:

- Propose a novel deepfake audio detection method leveraging dual channels of direct and reverberant sounds to differentiate between real and deepfake audio.
- Develop a new Sports Press Conference dataset for deepfake audio detection, offering longer and higher-quality recordings than existing datasets.
- Identify the optimal audio length for effective deepfake detection.

The paper will proceed as described below. Section II explains the related work. The mathematical preliminaries and our model design are described in Section III and IV. Experimental results and discussions are presented in Section V and VI. Finally, Section VII presents the conclusion.

II. RELATED WORK

A. DEEPFAKE AUDIO DETECTION W/ 1D WAVEFORM

Tak *et al.* [26] proposed the first application of RawNet2 [27] to prevent spoofing in Automatic Speaker Verification (ASV) systems. RawNet2 directly processes the raw audio input and uses temporal convolutional layers and an attention mechanism to capture both short-term and long-term features of the audio signal. Lavrentyeva *et al.* [28] demonstrated that integrating neural network-based models with handcrafted features can be effective. Additionally, ensemble models [29], [30] that combine 1D raw waveform approaches with 2D spectrograms have proven effective by leveraging the temporal information captured by raw waveforms and the frequency-domain characteristics from spectrograms, thereby improving robustness and detection accuracy. This model framework can be represented as shown in Figure 2(a-i).

B. DEEPFAKE AUDIO DETECTION W/ 2D SPECTROGRAM

Figure 2(a-ii) shows the deepfake audio detection framework that utilizes a 2D spectrogram as the single-channel input. Hamza *et al.* [9] stated that MFCC is useful for training both machine learning (ML) and DL models. Similarly, Qais *et al.* [10] found that MFCC contains richer information than other features, such as spectral centroid and spectrograms, when analyzing sound waves under computational constraints. Wu *et al.* [11] introduced a self-attention-based fake span strategy for deepfake audio detection, partially using MFCC and LFCC features. Other researchers have proposed various types of network architectures, such as Spec-ResNet [12] and CNN-LSTM [13], to detect deepfake audio using MFCC features. Arif *et al.* [31] proposed an extended local ternary pattern (ELTP) combined with LFCC to train a deep bidirectional long short-term memory (DBiLSTM) network for robust deepfake audio classification.

Previous research on deepfake audio detection has primarily focused on analyzing and manipulating single-channel data from 1D raw waveforms or 2D spectrograms to extract deepfake audio features. However, these studies have relied solely on features derived from generative models

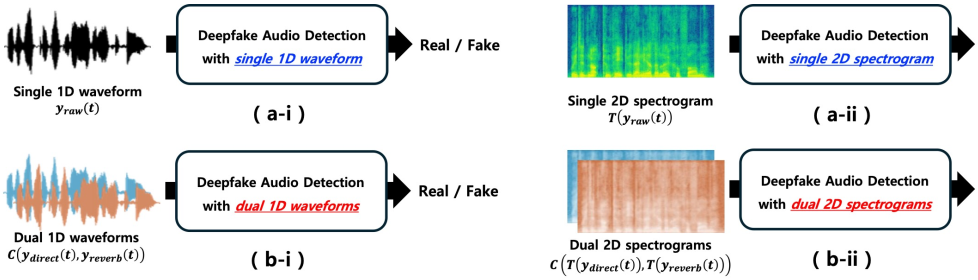
Deepfake Audio Detection with single 1D waveform
Real / Fake
Single 1D waveform $y_{raw}(t)$
(a-i)
Deepfake Audio Detection with single 2D spectrogram
Real / Fake
Single 2D spectrogram $T(y_{raw}(t))$
(a-ii)
Deepfake Audio Detection with dual 1D waveforms
Real / Fake
Dual 1D waveforms $C(y_{direct}(t), y_{reverb}(t))$
(b-i)
Deepfake Audio Detection with dual 2D spectrograms
Real / Fake
Dual 2D spectrograms $C\big(T(y_{direct}(t)), T(y_{reverb}(t))\big)$
(b-ii)

FIGURE 2: (a) Existing deepfake audio detection model framework using single-channel data and (b) proposed detection model framework using dual-channel data. (i) and (ii) show the models utilizing 1D waveform and 2D spectrogram, respectively. $T(\cdot)$ is a function that transfers a 1D waveform to the 2D spectrogram. $C(\cdot)$ is a function of combining direct and reverberant signals.

and overlooked real-world factors such as head movement, recording environments, and audio sources. To address this limitation, we propose a novel dual-channel deepfake audio detection model that incorporates real-world information by leveraging the relationship between direct and reverberant sounds. Unlike prior methods [9]–[13], [26], [28], [30], [31], which use ***single-channel data*** (1D raw waveforms or 2D spectrograms), as illustrated in Figure 2(a), the proposed framework utilizes ***dual-channel data*** (direct and reverberant 1D waveforms or 2D spectrograms) decomposed from single-channel data, as shown in Figure 2(b). The dual-channel data inherently captures real-world information that is challenging to infer from single-channel data. Our study demonstrates that the performance of deepfake audio detection models significantly improves by transitioning from single-channel to dual-channel training datasets.

III. MATHEMATICAL PRELIMINARIES

A. DIRECT AND REVERBERANT WAVEFORMS

Here, we first describe the concept of direct waveform $y_{direct}(t)$ and reverberant waveform $y_{reverb}(t)$. There waveforms are defined as the propagation of sound in an environment $h_{env}(t)$ where the sound from source $x_{source}(t)$ reaches the receiver like microphone either directly or through multiple reflections, as shown in Figure 1. The direct waveform $y_{direct}(t)$ represents the sound traveling directly from the source $x_{source}(t)$ to receiver without any reflections, is described as

$$y_{direct}(t) = x_{source}(t - \tau_0), \tag{1}$$

where τ_0 is the time delay corresponding to the direct path from source to receiver. The reverberant waveform $y_{reverb}(t)$ describes the reflected sound from various surfaces in the environment, is formulated as

$$\begin{aligned} y_{reverb}(t) &= (x_{source} * h_{env})(t) \\ &= \int_{-\infty}^{\infty} x_{source}(\tau) h_{env}(t - \tau) d\tau, \end{aligned} \tag{2}$$

where $*$ denotes convolution operation and $h_{env}(t)$ is the environment's impulse response function including reflections along various surfaces. The impulse response function $h_{env}(t)$ is usually defined by simple time delay modeling:

$$h_{env}(t) = \sum_{i=1}^{N} \omega_i \delta(t - \tau_i), \tag{3}$$

where N and ω_i denote the number of reflections and the attenuation factor for i^{th} reflection, respectively. $\delta(t - \tau_i)$ is the Dirac delta function with the time delay τ_i. Using Eq. 3, the reverberant waveform $y_{reverb}(t)$ can be rewritten as

$$\begin{aligned} y_{reverb}(t) &= \int_{-\infty}^{\infty} x_{source}(\tau) \sum_{i=1}^{N} \omega_i \delta(t - \tau_i - \tau) d\tau \\ &= \sum_{i=1}^{N} \omega_i \int_{-\infty}^{\infty} x_{source}(\tau) \delta(t - \tau_i - \tau) d\tau \\ &= \sum_{i=1}^{N} \omega_i x_{source}(t - \tau_i) = \sum_{i=1}^{N} \tilde{y}^{i}_{reverb}(t), \end{aligned} \tag{4}$$

where $\tilde{y}^{i}_{reverb}(t) = \omega_i x_{source}(t - \tau_i)$ denotes i^{th} reverberant sound reflected by the source sound $x_{source}(t)$ scaled by ω_i with the time delay τ_i.

Therefore, the raw waveform $y_{raw}(t)$ is described as the sum of direct waveform $y_{direct}(t)$ and reverberant waveform $y_{reverb}(t)$, as follows

$$\begin{aligned} y_{raw}(t) &= y_{direct}(t) + y_{reverb}(t) \\ &= y_{direct}(t) + \sum_{i=1}^{N} \tilde{y}^{i}_{reverb}(t) \\ &= x_{source}(t - \tau_0) + \sum_{i=1}^{N} \omega_i x_{source}(t - \tau_i). \end{aligned} \tag{5}$$

As formulated in Eq. 5, the raw waveform $y_{raw}(t)$ can be defined as a combination of the source waveform $x_{source}(t - \tau_0)$ and the weighted sum of the delayed source waveform $\omega_i x_{source}(t - \tau_i)$.

B. RELATIONSHIP BETWEEN REAL AND FAKE SAMPLES AND THE DIRECT-TO-REVERB RATIO

In the previous Section III-A, the direct $y_{direct}(t)$, the reverberant $y_{reverb}(t)$, and the raw waveforms $y_{raw}(t)$ were described. The raw waveform $y_{raw}(t)$ consists of the direct waveform

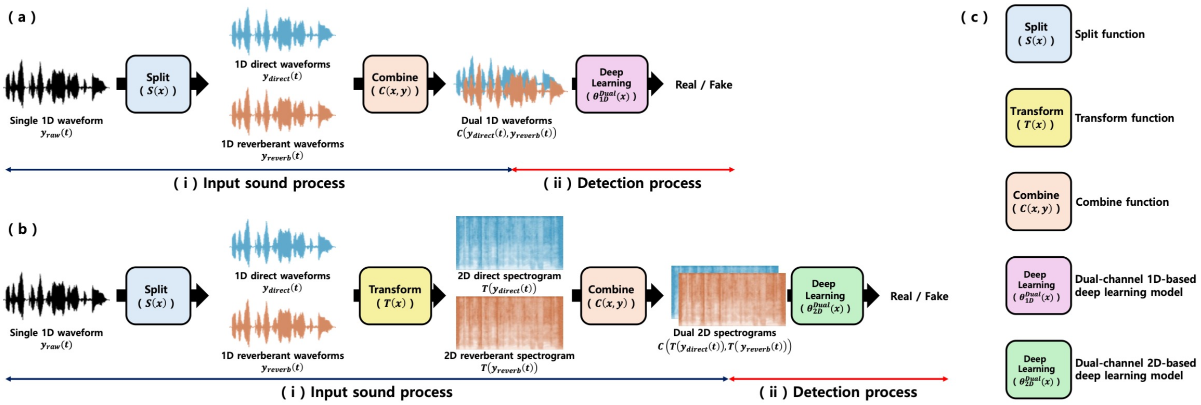
(a)
Single 1D waveform $y_{raw}(t)$
Split ($S(x)$)
1D direct waveforms $y_{direct}(t)$
1D reverberant waveforms $y_{reverb}(t)$
Combine ($C(x,y)$)
Dual 1D waveforms $C(y_{direct}(t), y_{reverb}(t))$
Deep Learning ($\theta_{1D}^{Dual}(x)$)
Real / Fake
(i) Input sound process
(ii) Detection process
(b)
Single 1D waveform $y_{raw}(t)$
Split ($S(x)$)
1D direct waveforms $y_{direct}(t)$
1D reverberant waveforms $y_{reverb}(t)$
Transform ($T(x)$)
2D direct spectrogram $T(y_{direct}(t))$
2D reverberant spectrogram $T(y_{reverb}(t))$
Combine ($C(x,y)$)
Dual 2D spectrograms $C(T(y_{direct}(t)), T(y_{reverb}(t)))$
Deep Learning ($\theta_{2D}^{Dual}(x)$)
Real / Fake
(i) Input sound process
(ii) Detection process
(c)
Split ($S(x)$) Split function
Transform ($T(x)$) Transform function
Combine ($C(x,y)$) Combine function
Deep Learning ($\theta_{1D}^{Dual}(x)$) Dual-channel 1D-based deep learning model
Deep Learning ($\theta_{2D}^{Dual}(x)$) Dual-channel 2D-based deep learning model

FIGURE 3: Overview of pipeline consisting of (i) input sound process and (ii) detection process. (a) and (b) show proposed dual-channel detection models using 1D waveform and 2D spectrogram, respectively. (c) Function modules used in (a) and (b).

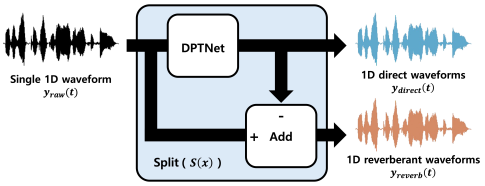
Single 1D waveform $y_{raw}(t)$
DPTNet
-
+ Add
Split ($S(x)$)
1D direct waveforms $y_{direct}(t)$
1D reverberant waveforms $y_{reverb}(t)$

FIGURE 4: Split function $\mathcal{S}(x)$ using DPTNet [32].

$y_{direct}(t)$ and the reverberant waveform $y_{reverb}(t)$, but the existing deepfake audio detection models introduced in Section II only use the 1D raw waveform $y_{raw}(t)$ or the 2D spectrogram STFT$(y_{raw}(t))$, obtained by applying the short-time Fourier transform (STFT) to the 1D raw waveform $y_{raw}(t)$. However, the important aspect is that information about the real environment is contained in the reverberant sound $y_{reverb}(t)$, to which the environment's impulse response $h_{env}(t)$ is applied as Eqs. 2 and 4.

The effect of the impulse response $h_{env}(t)$ is evident in the direct-to-reverberant ratio (DRR), defined as the ratio of the energy of the direct speech E_{direct} to the reverberant speech E_{reverb}, as follows

$$\mathrm{DRR} = 10 \log_{10}\left(\frac{E_{direct}}{E_{reverb}}\right), \tag{6}$$

where

$$E_{direct} = \sum_{t=1}^{T} |y_{direct}(t)|^2, \quad E_{reverb} = \sum_{t=1}^{T} |y_{reverb}(t)|^2.$$

T denotes the number of speech frames. For several datasets used in this study (more details are provided in Table 2), DRR scores were calculated in Table 1. Interestingly, the mean μ of DRR computed from fake samples is larger (while the standard deviation σ is smaller) than that from real samples. This clue suggests that fake samples generated by deepfake audio models do not account for various environmental factors, such as spatial location, volume, and the medium present in the reverberant sound $y_{reverb}(t)$. In other words, deepfake audio models fail to capture the characteristics of the reverberant waveform $y_{reverb}(t)$, which involves diverse environmental entities, when generating fake audio samples from the trained audio distribution.

TABLE 1: DRR scores depending on real and fake samples.

DRR	Public Datasets		Self-Collected
($\mu \pm \sigma$) [dB]	(a) ASVspoof2019 [33]	(b) FakeAVCeleb [34]	(c) SPC (V-A2)
Real	11.18 ± 22.55	13.94 ± 15.56	14.90 ± 16.08
Fake	$\mathbf{15.24 \pm 21.53}$	$\mathbf{22.97 \pm 11.53}$	$\mathbf{19.11 \pm 12.51}$

In this study, based on the limited characteristics of existing deepfake audio models, we propose a novel deepfake audio detection model that uses dual-channel data, including the direct waveform $y_{direct}(t)$ and the reverberant waveform $y_{reverb}(t)$ rather than a single-channel data such as the raw waveform $y_{raw}(t)$. For simplicity, we denote $y_{raw} := y_{raw}(t)$, $y_{direct} := y_{direct}(t)$, and $y_{reverb} := y_{reverb}(t)$.

IV. MODEL DESIGN

When a deepfake audio detection model detects audio modulation, this study proposes a pipeline, as shown in Figure 3, to efficiently leverage dual-channel waveforms, including the direct waveform y_{direct} and the reverberant waveform y_{reverb}. There are two types of proposed pipelines: a 1D waveform-based pipeline in Figure 3(a) and a 2D spectrogram-based pipeline in Figure 3(b). Additionally, each pipeline consists of an input sound process and a detection process, as shown in Figure 3(i)(ii). The following sections introduce details, including waveform splitting (see Section IV-A), feature transformation & combination (see Section IV-B), and model architectures (see Section IV-C).

A. WAVEFORM SPLITTING

In this study, the key idea is to leverage dual-channel signals, including a direct waveform y_{direct} and a reverberant waveform y_{reverb}, rather than the raw waveform y_{raw}. Therefore, the Dual Path Transformer Network (DPTNet) [32] is used to separate the direct waveform y_{direct} and the reverberant waveform y_{reverb} from the raw waveform y_{raw}. Figure 4 illustrates the working process of the split function $\mathcal{S}(x)$ using DPTNet [32]. Specifically, DPTNet [32] receives the 1D raw

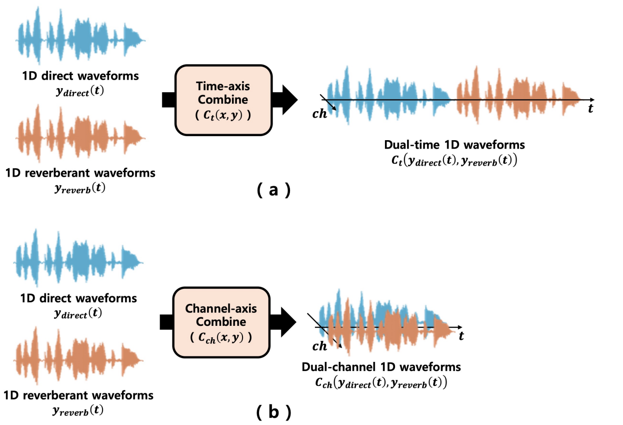

FIGURE 5: (a) C_t and (b) C_{ch} show along time-axis and channel-axis combination functions, respectively.

waveform y_{raw} as input and outputs only the direct waveform y_{direct} = DPTNet(y_{raw}). Subsequently, by subtracting the direct waveform y_{direct} from the raw waveform y_{raw}, the reverberant waveform $y_{reverb} = y_{raw} - y_{direct}$ is obtained.

B. FEATURE TRANSFORMATION & COMBINATION

To achieve the best performance in deepfake audio detection, it is necessary to verify the optimal combination of features generated from the waveform. In this study, three feature transformation methods were used: 1D wave, 2D MFCC, and 2D LFCC. Additionally, the time-axis combination C_t and the channel-axis combination C_{ch} were utilized as feature combination methods. The three feature transformation methods were applied to all types of waveforms, including the 1D raw waveform y_{raw}, direct waveform y_{direct}, and reverberant waveform y_{reverb}. The two types of combinations were applied only to dual-channel data, not single-channel data. Details of the combination functions are illustrated in Figure 5. Specifically, the time-axis combination function C_t, shown in Figure 5(a), connects two waveforms along the time axis, doubling the length of the waveform. Figure 5(b) illustrates the channel-axis combination function C_{ch}, which stacks two waveforms along the channel axis while maintaining the audio length.

C. MODEL ARCHITECTURES

To intensively verify the impact of the proposed dual-channel approach, well-known and intuitive deepfake audio detection models, including 1D models such as WaveRNN [35], TSSD [36], and RawNet [26], as well as 2D models such as ShallowCNN [37] and LCNN [38], are used in the experiments.

Depending on the feature transformations, the data structure can be a 1D shape like a waveform or a 2D shape such as MFCC and LFCC. WaveRNN [35], TSSD [36], and RawNet [26] are used as detection models for 1D data. To handle 2D features such as MFCC and LFCC, MLP, ShallowCNN [37], and LCNN [38] are employed. However, there are limitations in utilizing the proposed dual-channel data applied to the channel-axis combination C_{ch}. In WaveRNN [35], it is challenging to handle dual-channel data because the RNN architecture cannot process data with a channel axis. Similarly, RawNet [26], designed for 1D channel waveforms, cannot process data in the proposed dual-channel format. Furthermore, LCNN [38] is unsuitable for dual-channel data due to the Max-Feature-Map layer. To overcome this limitation, a convolutional layer is inserted after the dual-channel input data to convert the dual-channel data into single-channel data.

TABLE 2: Dataset details on the size of the training / validation / test set, the number of real and fake samples, waveform length, and the types of synthesis algorithms.

Dataset	Public Datasets		Self-Collected
	(a) ASVspoof2019 [33]	(b) FakeAVCeleb [34]	(c) SPC (V-A2)
# of Train (Real / Fake)	20,304 (2,064 / 18,240)	800 (400 / 400)	1,282 (220 / 1,062)
# of Valid (Real / Fake)	5,076 (516 / 4,560)	100 (50 / 50)	119 (20 / 99)
# of Test (Real / Fake)	71,237 (7,355 / 63,882)	266 (50 / 216)	400 (25 / 375)
# of subject (Male / Female) Male details Female details	78 (33 / 45) Train:8 / Dev:4 / Test:21 Train:12 / Dev:6 / Test:27	500 (250 / 250) Train:191 / Val:29 / Test:30 Train:209 / Val:21 / Test:20	204 (181 / 23) Train:144 / Val:19 / Test:18 Train:19 / Val:2 / Test:2
Speech length ($\mu \pm \sigma$) [sec]	3.54 ± 1.42	6.53 ± 2.51	27.75 ± 8.06
Speech frames (T) ($\mu \pm \sigma$) [frames]	$56,640 \pm 22,720$	$104,480 \pm 40,160$	$444,000 \pm 128,960$
Synthesis algo.	17 types of TTS and VS [33]	[1], [39], [40], [41]	[39], [40], [42]

V. EXPERIMENTS

A. DATASET

For this study, three datasets were used: two public datasets, ASVspoof2019 [33] and FakeAVCeleb [34], and one self-collected dataset, the sports press conference (SPC) dataset [43]. Details about each dataset are described in Table 2.

1) Public Datasets

ASVspoof2019 [33] includes two tasks: Physical Access (PA) and Logical Access (LA). This study focuses on the LA dataset, which deals with spoofing using digitally generated speech from TTS and VC technologies. Real audio was recorded from a total of 78 human speakers, 33 male and 45 female. The recorded audio is divided into three parts: training (8 male, 12 female), development (4 male, 6 female), and test (21 male, 27 female). In these experiments, only the training and test sets were used. To generate fake audio datasets, 17 speech synthesis and voice conversion toolkits were used. Six of the audio generation methods are labeled as known attacks and applied to the audio in the training set to construct the training dataset (20,304 samples) and validation dataset (5,076 samples). The other 11 methods are considered unknown and, together with two of the known attacks, are used to generate the test dataset (71,237 samples).

FakeAVCeleb [34] is a dataset consisting of 500 real clips selected from approximately 7-second clips of YouTube videos from VoxCeleb2 [44]. The dataset includes celebrities of various ethnicities, such as Caucasian, Black, and South Asian. Each of the 500 real clips corresponds to a different individual among the 500 celebrities, containing 250 males and 250 females. To create fake clips including fake audio & real video ($A_F V_R$), real audio & fake video ($A_R V_F$), and fake audio & fake video ($A_F V_F$), various facial and audio synthesis

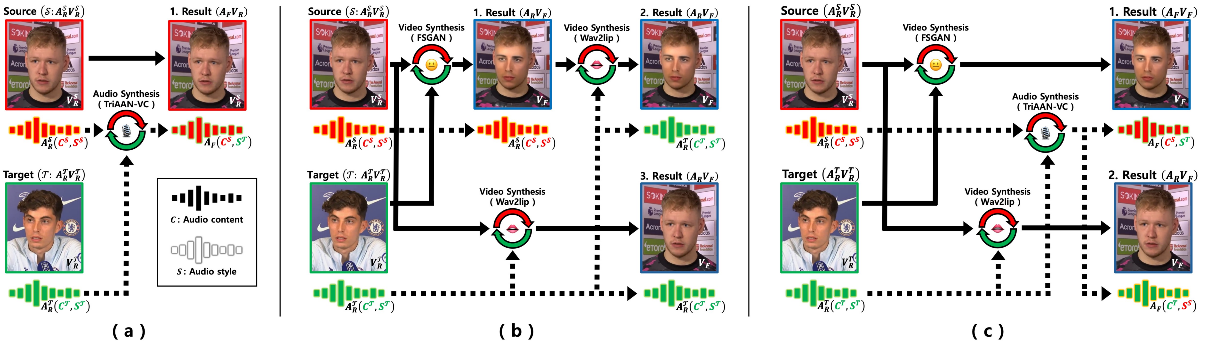

FIGURE 6: Pipeline to generate three types of deepfake data, (a) Fake Audio A_F & Real Video V_R ($A_F V_R$), (b) Real Audio A_R & Fake Video V_F ($A_R V_F$), and (c) Fake Audio A_F & Fake Video V_F ($A_F V_F$). Audio synthesis is performed by TriAAN-VC [42], and FSGAN [39] and Wav2lip [40] are used for video synthesis.

technologies were employed. In particular, Faceswap [1], FSGAN [39], and Wav2Lip [40] were used for face synthesis. Audio synthesis was performed using SV2TTS [41]. Although the FakeAVCeleb [34] dataset includes both video and audio tracks, this study utilized only the audio tracks. Specifically, a total of 1,000 audio files were used in $A_F V_R$, including 500 real and 500 fake audio files. The composition of the datasets is as follows: training (800 samples, with 191 male and 209 female), validation (100 samples, with 29 male and 21 female), and test (100 samples, with 30 male and 20 female). Moreover, 166 samples of $A_F V_F$ generated from combinations of test speakers were added to this test set.

2) Self-Collected Dataset

The SPC dataset [43] focuses on creating a longer dataset compared to the previous two datasets [33], [34]. The real clips are collected from press conference videos of sports stars posted on YouTube, with an average length of 30 seconds. A list of the clips used can be found at [43].

Real Dataset Collection. This dataset was collected from a total of 204 individuals, 181 male and 23 female. The clips are divided into three parts: training (144 male, 19 female), validation (19 male, 2 female), and test (18 male, 2 female). Each video was selected based on specific criteria:

1) Single speaker of English.
2) High-quality video to ensure clear face recognition.
3) Approximately 30-second in length.

A total of 265 clips of interviews with 204 individuals (primarily sports stars) were collected from YouTube videos as the real dataset, satisfying the above criteria.

Deepfake Dataset Generation. To generate deepfake clips from real clips, several deepfake-related methods were used. In particular, FSGAN [39] and Wav2Lip [40] were employed for video synthesis, such as face swap and lip generation, and TriAAN-VC [42] was used for deepfake audio.

Deepfake video V_F and/or deepfake audio A_F are generated by mixing source clips $\mathcal{S}$ containing real video $V_R^{\mathcal{S}}$ and real audio $A_R^{\mathcal{S}}$ with target clips $\mathcal{T}$ including real video $V_R^{\mathcal{T}}$ and real audio $A_R^{\mathcal{T}}$. Figure 6 shows the pipeline for generating three types of deepfake samples: (a) Fake audio A_F & real video V_R ($A_F V_R$), (b) real audio A_R & fake video V_F ($A_R V_F$), and (c) fake audio A_F & fake video V_F ($A_F V_F$). Details are below.

TABLE 3: Combination of Deepfake Dataset Generation.

(a) $A_F V_R$	**Fake audio A_F**	**Real video V_R**
Figure 6(a)	TriAAN-VC($A_R^{\mathcal{S}}$,$A_R^{\mathcal{T}}$)	Source vidio $V_R^{\mathcal{S}}$

(b) $A_R V_F$	**Real audio A_R**	**Fake video V_F**
Figure 6(b-1)	Source audio $A_R^{\mathcal{S}}$	FSGAN($V_R^{\mathcal{S}}$, $V_R^{\mathcal{T}}$)
Figure 6(b-2)	Target audio $A_R^{\mathcal{T}}$	Wav2lip(FSGAN($V_R^{\mathcal{S}}$, $V_R^{\mathcal{T}}$), $A_R^{\mathcal{T}}$)
Figure 6(b-3)	Target audio $A_R^{\mathcal{T}}$	Wav2lip($V_R^{\mathcal{S}}$, $A_R^{\mathcal{T}}$)

(c) $A_F V_F$	**Fake audio A_F**	**Fake video V_F**
Figure 6(c-1)	TriAAN-VC($A_R^{\mathcal{S}}$, $A_R^{\mathcal{T}}$)	Wav2lip(FSGAN($V_R^{\mathcal{S}}$, $V_R^{\mathcal{T}}$), $A_R^{\mathcal{S}}$)
Figure 6(c-2)	TriAAN-VC($A_R^{\mathcal{T}}$, $A_R^{\mathcal{S}}$)	Wav2lip($V_R^{\mathcal{S}}$, $A_R^{\mathcal{T}}$)

Fake audio & real video ($A_F V_R$) is created by manipulating source audio $A_R^{\mathcal{S}}$ with target audio $A_R^{\mathcal{T}}$ while keeping source video $V_R^{\mathcal{S}}$ as real video V_R, as shown in Figure 6(a). Specifically, the fake audio A_F can be created by changing the source audio style $S^{\mathcal{S}} = S(A_R^{\mathcal{S}})$ to the target audio style $S^{\mathcal{T}} = S(A_R^{\mathcal{T}})$ while maintaining the source audio content $C^{\mathcal{S}} = C(A_R^{\mathcal{S}})$. The fake audio $A_F(C^{\mathcal{S}}, S^{\mathcal{T}})$ is generated by TriAAN-VC [42]. Further details are given in Table 3(a).

Real audio & fake video ($A_R V_F$) is generated similarly to the generation process of $A_F V_R$. That is, in $A_F V_R$, only the audio track is manipulated, while in $A_R V_F$, only the video track is changed. Figure 6(b) shows three types of pipelines, and a description of each pipeline is given in Table 3(b). Here, FSGAN(V^1, V^2) performs a face swap from V^1 to V^2, and Wav2lip(V, A) manipulates lip movements in video V to match audio A.

Fake audio & fake video ($A_F V_F$) involves manipulating both the source audio $A_R^{\mathcal{S}}$ and the source video $V_R^{\mathcal{S}}$ of a source clip $\mathcal{S}$ using a target clip $\mathcal{T}$. There are two pipelines to create $A_F V_F$ as shown in Figure 6(c) and Table 3(c). TriAAN-VC(A^1, A^2) generates a converted audio track $A(C^1, S^2)$ by mixing the audio content C^1 of audio A^1 with the audio style S^2 of another audio A^2.

For 265 real clips, all pipelines in Figure 6 were repeated five times, resulting in a total of 7,950 deepfake clips. Similar to FakeAVCeleb [34], only the audio track was used in this study. Among three types of deepfake data: $A_F V_R$, $A_R V_F$, and

TABLE 4: Model details with respect to Feature, Channel, Data Type (D-Type), the number of parameters (# of Param.), used memory (Mem.), and runtime. At runtime, (i) and (ii) denote the RTF of the input sound process in Figure 2(i) and the detection process in Figure 2(ii), respectively.

Model	Feature	Channel	D-Type	# of Param.	Mem. (MB)	Runtime (RTF)	
						(i)	(ii)
WaveRNN [35]	WAVE	Single	y_{raw}	7,971,208	33.13	0.20	0.05
		Dual	C_t	15,190,408	62.92		0.09
			C_{ch}	7,971,208	35.43		0.07
TSSD [36]	WAVE	Single	y_{raw}	348,497	207.22		0.07
		Dual	C_t	348,497	413.04		0.09
			C_{ch}	348,609	207.99		0.06
RawNet [26]	WAVE	Single	y_{raw}	17,620,392	169.89		0.02
		Dual	C_t	17,620,392	270.37		0.02
			C_{ch}	17,620,392	172.20		0.02
MLP	MFCC/ LFCC	Single	y_{raw}	2,319,081	9.43	0.20	0.01
		Dual	C_t	4,627,881	18.82		0.01
			C_{ch}	4,627,881	18.82		0.01
ShallowCNN [37]	MFCC/ LFCC	Single	y_{raw}	1,104,401	18.59		0.02
		Dual	C_t	2,087,441	36.76		0.02
			C_{ch}	1,104,913	18.74		0.02
LCNN [38]	MFCC/ LFCC	Single	y_{raw}	14,474,081	177.51		0.09
		Dual	C_t	20,372,321	319.90		0.14
			C_{ch}	14,476,481	177.67		0.09

TABLE 5: Quantitative comparison with respect to Feature, Channel, and Data Type (D-Type). (a) and (b) show results for ASVspoof2019 [33] and FakeAVCeleb [34], respectively. The highest score for (Model, Feature) is in **bold**. The proposed channel-axis combination C_{ch} outperforms other data types for both 1D wave and 2D MFCC and LFCC features.

(a) **ASVspoof2019** [33]

Model	Feature	Channel	D-Type	AUC(↑)	EER(↓)
WaveRNN [35]	WAVE	Single	y_{raw}	0.5000	50.00 %
		Dual	C_t	0.5000	50.00 %
			C_{ch}	0.5000	50.00 %
TSSD [36]	WAVE	Single	y_{raw}	0.8227	20.81 %
		Dual	C_t	0.8959	10.77 %
			C_{ch}	**0.9303**	**7.90 %**
RawNet [26]	WAVE	Single	y_{raw}	0.9560	5.42 %
		Dual	C_t	0.9420	7.19 %
			C_{ch}	**0.9593**	**5.20 %**
MLP	MFCC	Single	y_{raw}	**0.8333**	21.65 %
		Dual	C_t	0.8176	**20.52 %**
			C_{ch}	0.8246	20.62 %
	LFCC	Single	y_{raw}	0.7143	34.25 %
		Dual	C_t	0.8115	23.14 %
			C_{ch}	**0.8361**	**18.06 %**
ShallowCNN [37]	MFCC	Single	y_{raw}	0.8677	16.91 %
		Dual	C_t	0.8878	15.31 %
			C_{ch}	**0.9266**	**8.98 %**
	LFCC	Single	y_{raw}	0.8023	27.52 %
		Dual	C_t	0.8075	27.17 %
			C_{ch}	**0.8169**	**25.89 %**
LCNN [38]	MFCC	Single	y_{raw}	0.9173	11.81 %
		Dual	C_t	0.8954	15.81 %
			C_{ch}	**0.9187**	**10.86 %**
	LFCC	Single	y_{raw}	0.8241	25.24 %
		Dual	C_t	0.8278	25.07 %
			C_{ch}	**0.8317**	**24.43 %**

(b) **FakeAVCeleb** [34]

Model	Feature	Channel	D-Type	AUC(↑)	EER(↓)
WaveRNN [35]	WAVE	Single	y_{raw}	0.8250	21.74 %
		Dual	C_t	0.7996	22.28 %
			C_{ch}	**0.8691**	**15.12 %**
TSSD [36]	WAVE	Single	y_{raw}	0.9654	5.71 %
		Dual	C_t	0.9692	4.08 %
			C_{ch}	**0.9707**	**3.92 %**
RawNet [26]	WAVE	Single	y_{raw}	0.9245	9.53 %
		Dual	C_t	0.9399	**6.02 %**
			C_{ch}	**0.9860**	8.06 %
MLP	MFCC	Single	y_{raw}	0.9030	11.47 %
		Dual	C_t	0.9831	1.99 %
			C_{ch}	**0.9900**	**1.96 %**
	LFCC	Single	y_{raw}	0.8844	11.89 %
		Dual	C_t	**0.9800**	**3.85 %**
			C_{ch}	**0.9800**	**3.85 %**
ShallowCNN [37]	MFCC	Single	y_{raw}	0.9800	3.85 %
		Dual	C_t	0.9884	2.26 %
			C_{ch}	**0.9900**	**1.96 %**
	LFCC	Single	y_{raw}	0.9368	9.80 %
		Dual	C_t	**0.9907**	**1.82 %**
			C_{ch}	0.9800	3.85 %
LCNN [38]	MFCC	Single	y_{raw}	0.9731	3.90 %
		Dual	C_t	0.9931	1.37 %
			C_{ch}	**1.0000**	**0.00 %**
	LFCC	Single	y_{raw}	0.9445	5.95 %
		Dual	C_t	0.9275	11.31 %
			C_{ch}	**0.9954**	**0.92 %**

$A_F V_F$, $A_F V_R$ and $A_F V_F$ were used. Therefore, the dataset for deepfake audio detection consisted of 265 real audios and 1,536 fake audios, and split into training (1,282 samples), validation (119 samples), and test (400 samples). Details are described in Table 2.

B. SETUP

1) Hyper Parameters

As the objective function for all deepfake audio detection models, binary cross-entropy loss (BCE loss) is used:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\omega_{pos}\cdot y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where N represents the total number of samples, and y_i and $\hat{y}i$ are the true label and the predicted probability for the i^{th} sample, respectively. Additionally, to address the data imbalance problem between real and fake samples, the positive weight $\omega_{pos} = \frac{\text{number of positive samples}}{\text{number of negative samples}}$ is applied to the BCE loss. The hyperparameters used to train the detection models are detailed as follows: The batch size was set to 64, the number of epochs was 50, and the Adam optimizer was used. The initial learning rate and weight decay were set to 10^{-4}. The optimal model was selected as the model that achieved the minimum validation loss.

2) Pre-Processing

The audio length in the ASVspoof2019 [33] and FakeAVCeleb [34] datasets is approximately 6 seconds. However, some audio clips may be shorter or longer than 6 seconds. Therefore, when feeding a waveform to the detection model, a pre-processing routine is applied to extract a waveform with a fixed length of 6 seconds. The pre-processing steps are as follows: If the audio length is longer than 6 seconds, a starting point is randomly selected, and the waveform for the 6 seconds after that point is extracted. On the other hand, if the audio length is shorter than 6 seconds, it is repeated until the audio exceeds 6 seconds, and then a random starting point is selected to extract 6-second waveform segments. Using random waveform segments reduces the risk of overfitting and improves the generalization ability of the detection models. Additionally, standard preprocessing routines such as data normalization and resampling are performed.

3) Evaluation Metrics

Two quantitative metrics, equal error rate (EER) and area under the curve (AUC), were used to evaluate deepfake audio

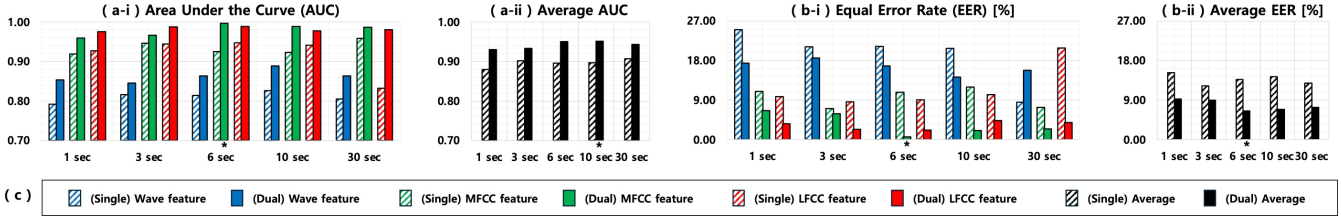

FIGURE 7: (a) AUC and (b) EER profiles for various audio lengths according to various features (waveform, MFCC, and LFCC) on the SPC dataset. The legend is defined in (c). Quantitative metrics for each feature are in (i), and (ii) shows the average scores for single-channel and dual-channel data. * denotes the best score. Models trained with 6-second segments of dual-channel MFCC feature achieved the best performance in both AUC and EER.

TABLE 6: Quantitative comparison with respect to Feature, Second (Sec.), Channel (Ch.), and Data Type (D-Type) on the SPC dataset. (a), (b), and (c) show results for WAVE, MFCC, and LFCC, respectively. Sec. denotes the length (in seconds) of the wave segment. The highest score for (Model, Feature) is in **bold**. Models trained with 6 or 10-second segments with the proposed dual-channel data type C_{ch} outperforms other data types for both 1D wave and 2D MFCC and LFCC features.

Feature				(a) WAVE	
Model	**Sec.**	**Ch.**	**D-Type**	**AUC(↑)**	**EER(↓)**
WaveRNN [35]	1	Single	y_{raw}	0.5339	48.23 %
		Dual	C_{ch}	0.6080	43.88 %
	3	Single	y_{raw}	0.5403	47.86 %
		Dual	C_{ch}	0.5950	44.75 %
	6	Single	y_{raw}	0.5277	48.54 %
		Dual	C_{ch}	0.6362	42.08 %
	10	Single	y_{raw}	0.5855	45.06 %
		Dual	C_{ch}	**0.6907**	**37.83 %**
	30	Single	y_{raw}	0.5547	47.08 %
		Dual	C_{ch}	0.6120	43.58 %
TSSD [36]	1	Single	y_{raw}	0.9057	15.87 %
		Dual	C_{ch}	0.9659	6.39 %
	3	Single	y_{raw}	0.9638	6.75 %
		Dual	C_{ch}	0.9751	4.74 %
	6	Single	y_{raw}	0.9860	2.73 %
		Dual	C_{ch}	0.9857	2.78 %
	10	Single	y_{raw}	0.9407	10.61 %
		Dual	C_{ch}	**0.9909**	**1.67 %**
	30	Single	y_{raw}	0.9400	10.71 %
		Dual	C_{ch}	0.9800	3.85 %
RawNet [26]	1	Single	y_{raw}	0.9375	10.90 %
		Dual	C_{ch}	0.9870	2.00 %
	3	Single	y_{raw}	0.9430	8.83 %
		Dual	C_{ch}	0.9655	6.36 %
	6	Single	y_{raw}	0.9299	12.30 %
		Dual	C_{ch}	0.9701	5.42 %
	10	Single	y_{raw}	0.9510	6.53 %
		Dual	C_{ch}	0.9831	3.28 %
	30	Single	y_{raw}	0.9200	13.79 %
		Dual	C_{ch}	**1.0000**	**0.00 %**

Feature				(b) MFCC	
Model	**Sec.**	**Ch.**	**D-Type**	**AUC(↑)**	**EER(↓)**
MLP	1	Single	y_{raw}	0.8781	17.56 %
		Dual	C_{ch}	0.9686	4.76 %
	3	Single	y_{raw}	0.9260	9.12 %
		Dual	C_{ch}	0.9229	13.34 %
	6	Single	y_{raw}	0.8447	23.23 %
		Dual	C_{ch}	**0.9902**	**1.87 %**
	10	Single	y_{raw}	0.8887	17.18 %
		Dual	C_{ch}	0.9661	6.35 %
	30	Single	y_{raw}	0.9333	10.84 %
		Dual	C_{ch}	0.9600	7.41 %
ShallowCNN [37]	1	Single	y_{raw}	0.9331	8.14 %
		Dual	C_{ch}	0.9414	10.41 %
	3	Single	y_{raw}	0.9387	8.90 %
		Dual	C_{ch}	0.9816	3.50 %
	6	Single	y_{raw}	0.9543	5.49 %
		Dual	C_{ch}	0.9997	0.06 %
	10	Single	y_{raw}	0.9384	8.10 %
		Dual	C_{ch}	**1.0000**	**0.00 %**
	30	Single	y_{raw}	0.9600	7.41 %
		Dual	C_{ch}	**1.0000**	**0.00 %**
LCNN [38]	1	Single	y_{raw}	0.9471	7.10 %
		Dual	C_{ch}	0.9692	4.69 %
	3	Single	y_{raw}	0.9732	3.14 %
		Dual	C_{ch}	0.9944	0.90 %
	6	Single	y_{raw}	0.9749	3.65 %
		Dual	C_{ch}	**1.0000**	**0.00 %**
	10	Single	y_{raw}	0.9407	10.61 %
		Dual	C_{ch}	**1.0000**	**0.00 %**
	30	Single	y_{raw}	0.9800	3.85 %
		Dual	C_{ch}	**1.0000**	**0.00 %**

Feature				(c) LFCC	
Model	**Sec.**	**Ch.**	**D-Type**	**AUC(↑)**	**EER(↓)**
MLP	1	Single	y_{raw}	0.9206	8.93 %
		Dual	C_{ch}	0.9798	2.72 %
	3	Single	y_{raw}	0.9172	13.18 %
		Dual	C_{ch}	0.9769	4.33 %
	6	Single	y_{raw}	0.9362	10.23 %
		Dual	C_{ch}	**0.9905**	**1.87 %**
	10	Single	y_{raw}	0.9195	13.33 %
		Dual	C_{ch}	0.9831	3.28 %
	30	Single	y_{raw}	0.9187	13.82 %
		Dual	C_{ch}	0.9800	3.85 %
ShallowCNN [37]	1	Single	y_{raw}	0.9517	6.68 %
		Dual	C_{ch}	0.9687	5.51 %
	3	Single	y_{raw}	0.9619	6.41 %
		Dual	C_{ch}	**0.9914**	**1.34 %**
	6	Single	y_{raw}	0.9744	4.48 %
		Dual	C_{ch}	0.9905	1.87 %
	10	Single	y_{raw}	0.9649	6.36 %
		Dual	C_{ch}	0.9831	3.28 %
	30	Single	y_{raw}	0.9800	3.85 %
		Dual	C_{ch}	0.9800	3.85 %
LCNN [38]	1	Single	y_{raw}	0.9093	13.75 %
		Dual	C_{ch}	0.9773	2.73 %
	3	Single	y_{raw}	0.9519	6.25 %
		Dual	C_{ch}	**0.9929**	**1.34 %**
	6	Single	y_{raw}	0.9296	12.34 %
		Dual	C_{ch}	0.9857	2.78 %
	10	Single	y_{raw}	0.9386	10.94 %
		Dual	C_{ch}	0.9661	6.35 %
	30	Single	y_{raw}	0.9773	4.34 %
		Dual	C_{ch}	0.9800	3.85 %

detection performance. EER represents the error rate when the false acceptance rate (FAR) and the false rejection rate (FRR) are equal, is described as

$$\mathrm{EER} = \mathrm{FAR}(\mathrm{threshold}^*) = \mathrm{FRR}(\mathrm{threshold}^*), \qquad (7)$$

where threshold* is the value when FAR and FRR are the same, FAR $= \frac{\text{Number of False Acceptances}}{\text{Number of Imposter Attempts}}$, and FRR $= \frac{\text{Number of False Rejections}}{\text{Number of Genuine Attempts}}$. AUC refers to the area under the receiver operating characteristic (ROC) curve. Here, the ROC curve is a graph drawn by calculating the true positive rate (TPR) and false positive rate (FPR) at various threshold. TPR and FPR are formulated as follows

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

where TP and TN represent true positive and true negative, respectively. FP is false positive, and FN is false negative. In addition, real-time factor (RTF) was computed to check the real-time capabilities, is formulated as

$$\mathrm{RTF} = \frac{\text{Processing time}}{\text{Audio duration}}.$$

C. RESULT

Table 4 shows the model details, including the number of parameters, memory used during processing, and runtime in terms of RTF. All models used in the experiments achieve an 0.3 RTF or less, which means that the entire process, comprising (i) input sound processing and (ii) detection processing, is performed within 0.3 seconds for 1 second of audio. In an IoT environment, user voice commands consist of short and concise words and last less than 3 seconds. Therefore, a typical voice commands can be inspected by the detection

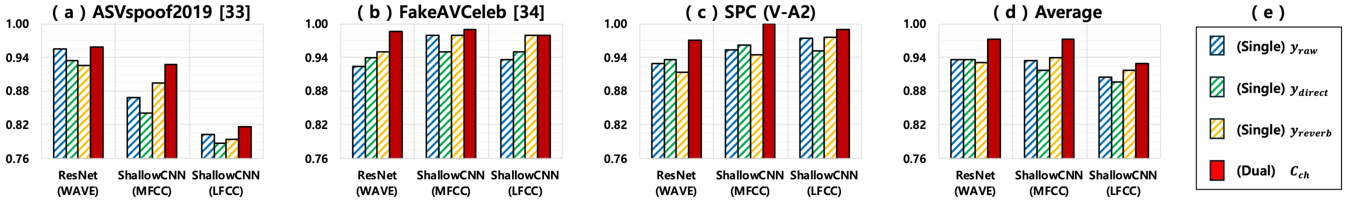

FIGURE 8: AUC profiles for (a) ASVspoof2019 [33], (b) FakeAVCeleb [34], (c) SPC (V-A2), and (d) average AUC profile for various single-channel data types (y_{raw}, y_{direct}, y_{reverb}) and the proposed dual-channel data type C_{ch}. The legend is defined in (e). Models trained with the proposed dual-channel data C_{ch} outperforms other single-channel data (y_{raw}, y_{direct}, y_{reverb}) in AUC.

TABLE 7: Quantitative comparison with respect to (i) ASVspoof2019 [33], (ii) FakeAVCeleb [34], and (c) SPC (V-A2). Models trained with each single dataset (a-c) and the entire dataset (d) are used to compute AUC and EER on the test set (i-iii). RawNet [26] and ShallowCNN [37] are used for WAVE and MFCC/LFCC features, respectively. The highest score for Model is in **bold**. Models trained with the entire dataset demonstrate superior generalization performance across datasets.

Training Dataset					(a) **ASVspoof2019** [33]		(b) **FakeAvCeleb** [34]		(c) **SPC**(V-A2)		(d) **All**	
Test Dataset	**Model**	**Feature**	**Channel**	**D-Type.**	**AUC(↑)**	**EER(↓)**	**AUC(↑)**	**EER(↓)**	**AUC(↑)**	**EER(↓)**	**AUC(↑)**	**EER(↓)**
(i) **ASVspoof2019** [33]	RawNet [26]	WAVE	Single	y_{raw}	0.9560	5.42 %	0.5117	49.41 %	0.5001	50.00 %	0.9068	13.58 %
			Dual	C_{ch}	**0.9593**	**5.20 %**	0.5000	50.00 %	0.5250	48.24 %	0.9413	8.25 %
	ShallowCNN [37]	MFCC	Single	y_{raw}	0.8677	16.91 %	0.5000	50.00 %	0.5000	50.00 %	0.8206	25.12 %
			Dual	C_{ch}	**0.9266**	**8.89 %**	0.4943	50.29 %	0.5000	50.00 %	0.8497	21.92 %
		LFCC	Single	y_{raw}	0.8023	27.52 %	0.3830	57.25 %	0.4203	54.58 %	0.7574	32.17 %
			Dual	C_{ch}	0.8169	**25.89 %**	0.5110	49.44 %	0.4757	51.25 %	**0.8187**	25.99 %
(ii) **FakeAvCeleb** [34]	RawNet [26]	WAVE	Single	y_{raw}	0.4838	50.82 %	0.9245	9.53 %	0.4631	51.98 %	**0.9977**	**0.46 %**
			Dual	C_{ch}	0.5000	50.00 %	0.9860	8.06 %	0.4865	50.19 %	0.9800	3.85 %
	ShallowCNN [37]	MFCC	Single	y_{raw}	0.3889	56.25 %	0.9800	3.85 %	0.5000	50.00 %	0.9792	4.04 %
			Dual	C_{ch}	0.5044	49.75 %	**0.9900**	**1.96 %**	0.5000	50.00 %	0.9761	2.76 %
		LFCC	Single	y_{raw}	0.4938	50.32 %	0.9368	9.80 %	0.5568	46.61 %	0.9368	9.80 %
			Dual	C_{ch}	0.5331	48.27 %	**0.9800**	**3.85 %**	0.5830	45.08 %	0.9469	5.92 %
(iii) **SPC** (V-A2)	RawNet [26]	WAVE	Single	y_{raw}	0.5000	50.00 %	0.5000	50.00 %	0.9299	12.30 %	**0.9973**	**0.53 %**
			Dual	C_{ch}	0.5093	49.45 %	0.5000	50.00 %	0.9701	5.42 %	0.9933	1.32 %
	ShallowCNN [37]	MFCC	Single	y_{raw}	0.5600	46.81 %	0.5000	50.00 %	0.9543	5.49 %	0.9773	4.34 %
			Dual	C_{ch}	0.5547	47.08 %	0.5000	50.00 %	**0.9997**	**0.06 %**	0.9720	3.91 %
		LFCC	Single	y_{raw}	0.5120	49.36 %	0.3693	58.06 %	0.9744	4.48 %	0.9613	3.99 %
			Dual	C_{ch}	0.5600	46.81 %	0.4827	50.92 %	0.9905	1.87 %	**0.9960**	**0.79 %**

model in under 1 second. In addition, since the model requires less than 1GB of memory, it is suitable for real devices.

Table 5 shows the performance of all models on (a) ASVspoof2019 [33] and (b) FakeAVCeleb [34] datasets. The detection models leveraging the proposed dual-channel data outperform most single-channel based models except MLP trained with MFCC feature in Table 5(a). Specifically, in an environment utilizing the 1D WAVE feature, RawNet [26] using dual-channel waveform combined along the channel-axis C_{ch} achieves the highest AUC score compared to other wave feature-based models on both datasets. In EER metrics, RawNet [26] and TSSD [36] trained with the proposed dual-channel data show the best performance on ASVspoof2019 [33] and FakeAVCeleb [34], respectively. When applying 2D features, ShallowCNN [37] with MFCC-based dual-channel data is superior to others in ASVspoof2019 [33], but LCNN [38] using MFCC-based dual-channel data shows the best performance in FakeAVCeleb [34].

From dataset perspective, RawNet [26] trained with proposed 1D waveform-based dual-channel data achieves the highest score on ASVspoof2019 [33], while LCNN [38] with proposed dual-channel data on 2D LFCC shows the best performance on FakeAVCeleb [34].

VI. DISCUSSION

A. PERFORMANCE ACCORDING TO COMBINE FUNCTIONS

The major contribution of this study is leveraging dual-channel data consisting of the direct waveform y_{direct} and the reverberant waveform y_{reverb}. To feed the dual-channel data to the detection models, the direct waveform y_{direct} and the reverberant waveform y_{reverb} must be combined. In Section IV-B, two types of combining mechanisms are introduced: time-axis combination C_t and channel-axis combination C_{ch}, as shown in Figure 5. To validate the effect of these combining mechanisms, all detection models were trained using dual-channel data with both the time-axis combination C_t and the channel-axis combination C_{ch}, respectively. In Table 5, the dual-channel part shows the quantitative metrics according to the time-axis combination C_t and the channel-axis combination C_{ch}. For all 1D wave feature-based models, leveraging dual-channel data with the channel-axis combination C_{ch} outperforms using the time-axis combination C_t. This trend, observed in models based on 1D wave features, is also seen in models using 2D MFCC and LFCC features, excluding ShallowCNN [37]. Through this experiment, it was confirmed that the channel-axis combination C_{ch} is more suitable than the time-axis combination C_t for leveraging dual-channel data containing the direct waveform y_{direct} and the reverberant waveform y_{reverb}.

B. IMPACT OF WAVEFORM FEATURES & LENGTH

To verify changes in performance depending on waveform length, our group used the SPC self-collected dataset (average length 30-second), which has a longer waveform length than public datasets (average length 6-second) such as ASVspoof2019 [33] and FakeAVCeleb [34]. Wave samples from the SPC dataset were divided into various lengths, such as 1, 3, 6, 10, and 30 seconds, and were used to train the deepfake audio detection models. Table 6 shows the quantitative metrics for various waveform features and lengths, comparing single-channel models with the proposed dual-channel models. Generally, the detection models trained with the proposed dual-channel scheme (see solid bars in Figure 7) show better performance than the single-channel models (see dashed bars in Figure 7).

Figure 7(i) shows the performance for each feature type. In particular, 2D features including MFCC (see green bars in Figure 7) and LFCC (see red bars in Figure 7) are superior to the 1D wave feature (see blue bars in Figure 7). Among MFCC and LFCC features, using the MFCC feature at 6-second segment shows the best performance, as indicated by * in Figure 7(i). From the perspective of waveform lengths, using 10-second dual-channel segments achieves the best AUC score compared to other lengths, as shown in Figure 7(a-ii), while using 6-second dual-channel segments shows the best results in average EER in Figure 7(b-ii).

Through this experiment, our finding is that 2D features such as MFCC and LFCC are superior to 1D feature like waveform. Waveform segments that are too short or long are not suitable for extracting deepfake features and determining whether the audio is a deepfake. Using 6- or 10-second audio segments is more efficient for developing deepfake audio detection models. Public datasets such as ASVspoof2019 [33] and FakeAVCeleb [34] already satisfy this audio length.

C. IMPACT OF DATA TYPES

In Section III, we confirmed there are differences in the perception of direct y_{direct} and reverberant waveforms y_{reverb} between real and fake audio, as shown in the DRR scores in Table 1. Based on the insight, models using dual-channel data $C_{ch}(y_{direct}, y_{reverb})$ is proposed and shows better performance than models using raw data $y_{raw} = y_{direct} + y_{reverb}$. Here, we conducted a comparative study on models using various data types, including single-channel data (y_{raw}, y_{direct}, y_{reverb}) and dual-channel data (C_{ch}), to elucidate the impact of each data type. Figure 8 shows the AUC profiles for various single-channel data types and the dual-channel data type. It is clear that using the dual-channel data type outperforms the models trained with single-channel date types in the entire experimental environment. Among the single-channel data types, the models trained with data containing reverberant waveform y_{reverb}, such as raw waveform y_{raw} or reverberant waveform y_{reverb}, outperformed the models trained exclusively with direct waveform y_{direct}, as shown in Figure 8(d). From the experiment, it can be concluded that leveraging reverberant waveform y_{reverb} is important for detecting deepfake audio.

D. GENERALIZABILITY ACROSS DATASETS

Since each dataset used in the experiment has distinct characteristics, it is difficult to infer a model trained on a given dataset to other datasets. Table 7 shows the results of cross-validation for pairs of datasets and trained models. When the training and test data belong to the same dataset, the model generally achieves the best scores, as shown in Table 7(a-i), (b-ii), and (c-iii). However, when inference is performed on different datasets, performance deteriorates significantly. That is, it is difficult for a model trained with a single dataset to satisfy well-generalizability to other datasets. However, as shown in Table 7, the model trained using the entire datasets achieves similar performance to the model trained with individual datasets that matches the test set. These results demonstrate that training the model on the entire collected dataset improves generalizes across datasets.

VII. CONCLUSION

The study proposed a deepfake audio detection model leveraging dual-channel data. In a conversation situation or recording studio, waveforms captured by receivers like microphones include various environmental factors, such as the subject's behavior and the recording environment. Although these environmental factors cause differences between direct and reverberant waveforms, deepfake audio generation models do not account for these changes. Therefore, we proposed novel dual-channel detection models utilizing direct and reverberant waveforms rather than single-channel raw waveforms. By leveraging the proposed dual-channel data, most deepfake audio detection models using wave, MFCC, and LFCC features demonstrated improved performance, such as higher AUC and lower ERR, compared to models using single-channel data. Additionally, to verify the impact of various wave lengths, our group collected the SPC dataset with a longer wave length of around 30 seconds, compared to the public datasets of around 6 seconds. In experiments with different wave lengths, using wave segments of approximately 6 or 10 seconds showed the best performance among wave segments with lengths of 1, 3, 6, 10, and 30 seconds.

Overall, our findings indicate that using dual-channel audio data significantly improves the performance of deepfake audio detection and highlights the importance of capturing both direct and reverberant sound characteristics for robust and reliable detection. As future work, we will develop a deepfake detection model leveraging multi-modality, including the proposed dual-channel data and video information such as lip synchronization [45] and lip reading [46]. To enhance generalization performance, continuous collection of diverse datasets and the AI-based data augmentation such as environmental noise generation [47], [48] and audio source alteration [49] will be used as effective strategies.

REFERENCES

[1] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[2] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[4] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang *et al.*, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *arXiv preprint arXiv:2005.05535*, 2020.

[5] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[6] T. Halperin, A. Ephrat, and S. Peleg, "Dynamic temporal alignment of speech to lips," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3980–3984.

[7] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 439–447.

[8] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[9] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[10] A. Qais, A. Rastogi, A. Saxena, A. Rana, and D. Sinha, "Deepfake audio detection with neural networks using audio features," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*. IEEE, 2022, pp. 1–6.

[11] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9236–9240.

[12] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[13] I. Altalahin, S. AlZu'bi, A. Alqudah, and A. Mughaid, "Unmasking the truth: A deep learning approach to detecting deepfake audio through mfcc features," in *2023 International Conference on Information Technology (ICIT)*. IEEE, 2023, pp. 511–518.

[14] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[15] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 667–684.

[16] A. Khormali and J.-S. Yuan, "Add: Attention-based deepfake detection approach," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 49, 2021.

[17] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[18] S. Asha, P. Vinod, and V. G. Menon, "A defensive attention mechanism to detect deepfake content across multiple modalities," *Multimedia Systems*, vol. 30, no. 1, p. 56, 2024.

[19] A. Das, S. Das, and A. Dantcheva, "Demystifying attention mechanisms for deepfake detection," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–7.

[20] B. Taşcı, "Deep-learning-based approach for iot attack and malware detection." *Applied Sciences (2076-3417)*, vol. 14, no. 18, 2024.

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[22] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[23] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[24] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[25] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[26] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[27] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech*, pp. 3583–3587, 2020.

[28] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[29] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," *arXiv preprint arXiv:1904.04589*, 2019.

[30] S. Shukla, J. Prakash, and R. S. Guntur, "Replay attack detection with raw audio waves and deep learning framework," in *2019 International Conference on Data Science and Engineering (ICDSE)*, 2019, pp. 66–70.

[31] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.

[32] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[33] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[34] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[35] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[36] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[37] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, """ hello? who am i talking to?" a shallow cnn approach for human vs. bot speech classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2577–2581.

[38] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[39] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.

[40] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484–492.

[41] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[42] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conver-

sion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[43] G. Lee, J. Lee, M. Jung, and Y. Han, "Dual-channel deepfake audio detection: Leveraging direct and reverberant waveforms," https://github.com/gunwoo5034/Dual-Channel-Audio-Deepfake-Detection, Github, 2024.

[44] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[45] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "Lip sync matters: A novel multimodal forgery detector," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1885–1892.

[46] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[47] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[48] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[49] Supertone, Inc., "Ai realistic voice rendering (air)," https://www.supertone.ai/air, Supertone, 2024.

GUNWOO LEE received the B.S. degrees in electronic engineering from the University of Soongsil in 2025, and is currently pursuing a Master's degree under the supervision of Professor Yoseob Han at School of Electronic Engineering, Soongsil University, Seoul, South Korea. His current research interests include deepfake audio detection, medical image processing.

JUNGMIN LEE received the B.S. degrees in electronic engineering from the University of Soongsil in 2025, and is currently pursuing a Master's degree under the supervision of Professor Yoseob Han at School of Electronic Engineering, Soongsil University, Seoul, South Korea. His current research interests include deepfake audio detection, medical image processing.

MINKYO JUNG received the B.S. degrees in electronic engineering from the University of Soongsil in 2025, and is currently pursuing a Master's degree under the supervision of Professor Yoseob Han at School of Electronic Engineering, Soongsil University, Seoul, South Korea. His current research interests include deepfake audio detection, medical image processing.

JOSEPH LEE received the B.S. degrees in Statistics and Actuarial Science from the University of Soongsil in 2025. His current research interests include deepfake audio detection, medical image processing.

KIHUN HONG (Member, IEEE) received the B.S. degree in electronics engineering and the M.S. and Ph.D. degrees in information and telecommunication engineering from Soongsil University, Seoul, South Korea, in 2000, 2002, and 2006, respectively. From 2006 to 2007, he was a Postdoctoral Researcher with the University of California at Davis, USA. He was a Principal Engineer with Secui, South Korea, from 2008 to 2022, a network security company. He is currently an Assistant Professor with the School of Electronic Engineering, Soongsil University. His current research interests include deepfake detection, zero trust architecture, and industrial cybersecurity.

SOUHWAN JUNG (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, in 1985 and 1987, respectively, and the Ph.D. degree from the University of Washington, Seattle, USA, in 1996. From 1996 to 1997, he was a Senior Software Engineer with Stellar One Corporation, Bellevue, USA. In 1997, he joined the School of Electronic Engineering, Soongsil University, Seoul, South Korea, where he is currently a Professor. He has spent about 25 years on the area of information security. He served as a government R&D program director of information security during 2009–2010. He also served as the president of KIISC in 2020. Currently, he leads an AI security research center (AISRC) funded by government since 2020. AISRC focuses on developing robust and trustworthy AI systems. Currently, seven faculty members from three universities and six companies are participating in the consortium. His research interest includes deepvoice detection and other AI security issues.

YOSEOB HAN received the B.S. degrees in biomedical engineering from Kyung Hee University, in 2013. He received the M.S. and Ph.D degrees in bio and brain engineering from KAIST, in 2015 and 2019. From 2019 to 2020, He was a Postdoctoral Researcher in applied mathematics and plasma physics (T-5) and applied modern physics (P-21) groups at the Los Alamos National Laboratory, USA. From 2020 to 2022, he was a Postdoctoral Researcher in Department of Radiology and Deep learning approach for medical image reconstruction and diagnosis at Harvard Medical School and Massachusetts General Hospital, USA. In 2022, he worked Wecover Platform., South Korea, for Technical Lead and Parametric insurance and Pension service using blockchain. In 2023, he joined the School of Electronic Engineering, Soongsil University, Seoul, South Korea, where he currently serves as an Assistant Professor. His current research interests include Natural, Medical, Industrial image processing, and Speech signal processing.