Discriminative Feature Decoupling Enhancement for Speech Forgery Detection

Yijun Bei[1], **Xing Zhou**[1,2,3], **Erteng Liu**[1,2,3], **Yang Gao**[1,3], **Sen Lin**[4], **Kewei Gao**[1,2,3] and **Zunlei Feng**[1,2,3*]
[1]School of Software Technology, Zhejiang University
[2]State Key Laboratory of Blockchain and Security, Zhejiang University
[3]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[4]Ningbo Donghai Group Co., Ltd.
zunleifeng@zju.edu.cn

Abstract

The emergence of AIGC has brought attention to the issue of generating realistic deceptive content. While AIGC has the potential to revolutionize content creation, it also facilitates criminal activities. Specifically, the manipulation of speech has been exploited in tele-fraud and financial fraud schemes, posing a significant threat to societal security. Current deep learning-based methods for detecting forged speech extract mixed features from the original speech, which often contain redundant information. Moreover, these methods fail to consider the distinct characteristics of human voice-specific features and the diversity of background environmental sounds. This paper introduces a framework called Discriminative fEature dEcoupling enhanceMent (DEEM) for detecting speech forgery. Initially, the framework decouples the original speech into human voice features and background sound features. Subsequently, DEEM enhances voice-specific features through temporal dimension aggregation and improves continuity-related features in the background sound map via spectral-dimension aggregation. By employing the decoupling enhancement features, extensive experiments demonstrate that DEEM achieves an accuracy improvement of over 5% on FoR dataset compared to the state-of-the-art methods.

1 Introduction

With the rapid development of Artificial Intelligence in Generative Content (AIGC) techniques, the generated content has attained a remarkably realistic effect, capable of deceiving even human observers. AIGC is employed to generate textual, visual, auditory, and audiovisual content, thereby enhancing the efficiency of multimedia designers, facilitating educational purposes, and potentially revolutionizing the field of content generation.

Unfortunately, AIGC also presents opportunities for criminal exploitation. Unethical individuals harness the capabilities of AIGC to engage in criminal activities, including tele-fraud, financial fraud, and the dissemination of rumors. These actions pose significant threats to societal well-being and the overall fabric of humanity.

Numerous researchers are actively devoted to the advancement of forgery detection techniques, with a predominant emphasis on image [Agarwal and Verma, 2020; Koptyra and Ogiela, 2020; Guillaro *et al.*, 2023] and video [Afchar *et al.*, 2018; Zheng *et al.*, 2021] forgery detection. Comparatively, research on speech forgery detection remains scarce. Nonetheless, it is essential to recognize the significance of speed forgery detection techniques, particularly in combating tele-fraud and financial fraud, wherein solely synthetic speech is employed for criminal activities.

The existing methods for speech forgery detection can be categorized into two main groups: hand-crafted feature-based [Wu *et al.*, 2018; Alzantot *et al.*, 2019] and deep learning-based approaches [Tak *et al.*, 2021b; Tak *et al.*, 2021a]. The former involves extracting various speech features, such as cqtspec, logspec, and LFCC and then employing traditional classifiers to identify forgery features. However, the hand-crafted features designed by humans often fail to capture certain discriminative features, resulting in poor identification performance. On the other hand, the latter approach utilizes the original waveform or its variations as input and employs deep networks to extract critical discriminative forgery features, guided by labeled data. Compared to hand-crafted feature-based methods, deep learning-based methods achieve superior performance by effectively extracting supervised discriminative forgery features. Nevertheless, mixed features extracted by deep models still tend to contain redundant information for speech forgery detection purposes.

The speech typically consists of the prominent human voice and the accompanying environmental sounds. As a result, there are two primary sources of counterfeiting: the fabrication of human voices and the manipulation of background sounds. These factors lead to three categories of forged speech: speech with counterfeit human voice and counterfeit background sound, speech with counterfeit human voice and authentic background sound, and speech with genuine human voice and counterfeit background sound. Such forged speech is typically created using either AI models or human-based synthesis techniques. The synthesized methods encompass speech fragment merging, waveform concatenation synthesis, speech element editing, and others.

*Corresponding author

Based on the aforementioned synthesis technique, it can be observed that most forged speeches contain a significant amount of genuine components, which can negatively affect identification performance. For instance, when counterfeit human voices are combined with authentic background sounds, the features present in the authentic background sound can interfere with the identification of the counterfeit human voice feature.

Furthermore, human voices and background environmental sounds possess distinct characteristics. The former exhibits specific traits such as rate, timbre, intonation, and tone, which are vital for identifying the unique characteristics of a specific individual. On the other hand, background environmental sounds tend to be diverse and noisy. They may also include mixed voices from other individuals, further complicating the accurate identification of the prominent human voice. Hence, the counterfeit identification of background environmental sounds should focus on continuity and other relevant factors.

In this paper, we propose a Discriminative fEature dEcoupling enhanceMent framework (DEEM) for speech forgery detection, based on the distinct characteristics of prominent human voices and background environmental sounds. DEEM employs the swapping decoupling strategy to separate the speech into two distinct feature maps: the human voice feature map and the background sound feature map. Subsequently, we apply temporal dimension aggregation on the human voice feature map to augment the voice-specific features, and spectral-dimension aggregation on the background sound feature map to enhance continuity-related features. Furthermore, the enhanced feature vectors are integrated into a fully-connected heterogeneous graph. To extract forgery features for speech detection, we employ a widely used heterogeneous graph attention network on the heterogeneous graph. The results of extensive experiments demonstrate that DEEM, by combining temporal-aggregated human voice features and spectral-aggregated background sound features, achieves more than 5% accuracy improvement on FoR dataset compared to SOTA methods.

Our contribution lies in decoupling speech into prominent human voice features and background environmental sound features using a swapping strategy. This decoupling effectively minimizes the interference caused by redundant features in speech forgery detection. Additionally, we introduce temporal-dimension and spectral-dimension aggregations to enhance the human voice-specific features and continuity-related background features, respectively. Extensive experiments demonstrate that the proposed framework, incorporating the decoupled and enhanced features, achieves state-of-the-art performance.

2 Related Work

In this section, we present a concise survey of two techniques that are highly relevant in the context of forgery audio detection and feature decoupling.

2.1 Forgery Audio Detection

Artificially generated deceptive speech, known as forged speech, poses significant risks and threats. To mitigate these risks and protect individuals' interests, voice anti-counterfeiting detection techniques have been developed to provide digital security. The primary strategies employed in speech forgery are speech synthesis (TTS), which generates speech from text, and voice conversion (VC), which imitates the target timbre to alter the sound of speech. These forgery techniques are commonly exploited in voice fraud and spoofing verification systems, thereby posing a significant threat to information security.

Speech synthesis technology employs text analysis and waveform generation to extract phoneme information and generate speech waveforms. Traditional techniques for speech synthesis include waveform concatenation methods such as PSOLA and unit selection systems based on HMM [Bigorgne *et al.*, 1993; Yoshimura, 2002]. Additionally, parameter-based methods are used. However, the quality of synthesized speech has significantly improved with the emergence of deep learning advancements, such as WaveNet [Oord *et al.*, 2016], Deep Voice [Arık *et al.*, 2017], and Tacotron [Wang *et al.*, 2017]. Voice conversion, on the other hand, aims to establish a mapping between the source speech and the desired speech of the target speaker. This process can be divided into two categories: parallel corpus-based and non-parallel corpus-based methods. In parallel corpus speech conversion, frame alignment and feature mapping techniques, such as dynamic time warping (DTW) [Helander *et al.*, 2008], are employed. Statistical modeling methods, including Gaussian mixture models [Aihara *et al.*, 2013], are also commonly utilized. Non-parallel corpus speech conversion poses greater complexity, incorporating techniques such as alignment based on nearest neighbor search [Sun *et al.*, 2016], as well as the application of neural networks for feature parameter mapping [Desai *et al.*, 2009]. Various vocoders are employed in speech synthesis, including STRAIGHT [Kaneko and Kameoka, 2018], WORLD [Morise *et al.*, 2016], and WaveNet [Oord *et al.*, 2016]. These vocoders play a significant role in the generation of high-quality synthesized speech.

Existing deep learning-based methods have commonly utilized mixed deep features for speech forgery detection. [Alzantot *et al.*, 2019] proposed a deep ResNet-based anti-counterfeiting scheme that combines multiple features. [Li *et al.*, 2021] learned multi-scale features based on Res2Net to enhance model generalization. [Wu *et al.*, 2018] demonstrated that LCNN is suitable for speech anti-counterfeiting as it retains core information and strengthens feature learning. The ASSERT system, introduced by [Lai *et al.*, 2019], fuses ResNet and SENet. [Li *et al.*, 2021] further improved model performance by integrating the squeeze-excitation module into Res2Net. In addition, the Transformer model has been applied and innovated in many fields [Chen *et al.*, 2024], including speech forgery detection [Zhang *et al.*, 2021c]. Raw audio is used as the input for end-to-end anti-spoofing models. [Tak *et al.*, 2021b] utilized SincNet to process raw audio and proposed an end-to-end anti-spoofing algorithm called RawNet2. Furthermore, [Tak *et al.*, 2021a] employed a graph attention network to model different subbands and periods and enhance the performance of the model. [Jung *et al.*, 2022] adopted the heterogeneous graph-based attention mechanism

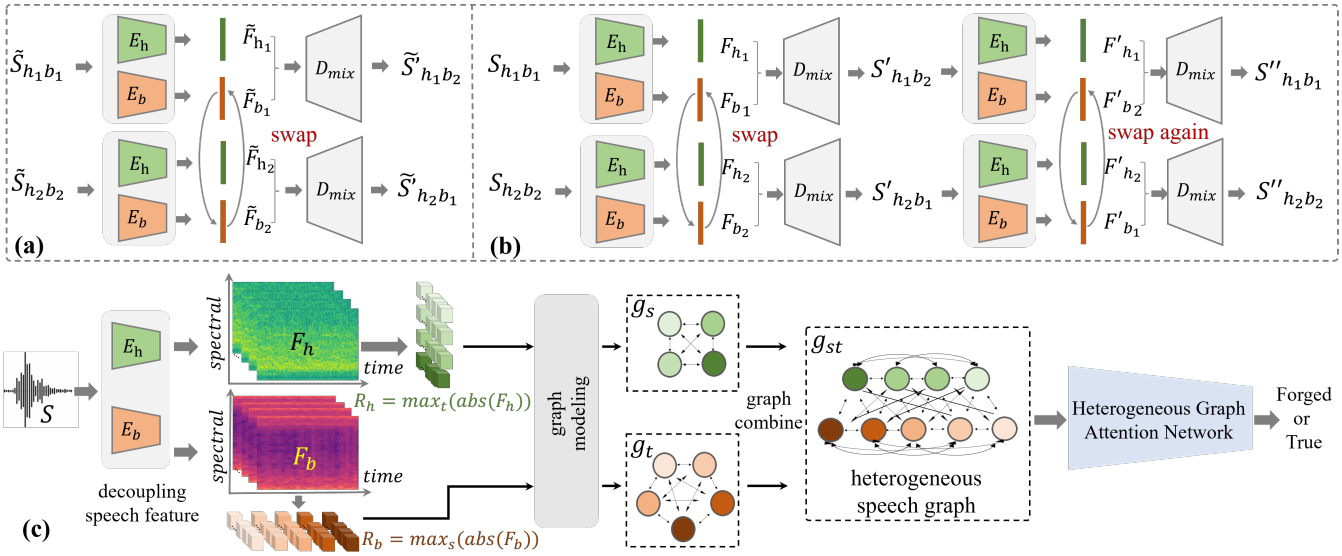

Figure 1: The Discriminative fEature dEcoupling enhanceMent (DEEM) framework for speech forgery detection. (a) The synthetic speeches $\widetilde{S}_{h_1b_2}$ and $\widetilde{S}_{h_2b_1}$ guide the encoders E_h and E_b in decoupling the input synthetic speeches $\widetilde{S}_{h_1b_1}$ and $\widetilde{S}_{h_2b_2}$ into distinct human voice features ($\widetilde{F}_{h_1}$ and $\widetilde{F}_{h_2}$) and background environment features ($\widetilde{F}_{b_1}$ and $\widetilde{F}_{b_2}$) using a swapping operation. (b) Self-supervised reconstruction is employed between the decoded speech $\widetilde{S}''_{h_1b_1}$ / $\widetilde{S}''_{h_2b_2}$ and the real speech $\widetilde{S}_{h_1b_1}$ / $\widetilde{S}_{h_2b_2}$ to decouple the input real speeches through a dual-swapping operation. (c) By employing pretrained encoders E_h and E_b, the input speech S is decoupled into a human voice feature map F_h and a background environment feature map F_b. Temporal- and spectral-dimension aggregations are subsequently applied to enhance human voice-specific features and continuity-related background features, respectively. These enhanced feature vectors are then used to model a fully connected heterogeneous speech graph $\mathcal{G}_{st}$. Finally, the Heterogeneous Graph Attention Network is employed for identifying speech forgery with the graph $\mathcal{G}_{st}$ as input.

for speech forgery detection through integrating frequency domain and spectral features as input.

However, the commonly adopted mixed features extracted by deep models often contain redundant information for speech forgery detection.

2.2 Feature Decoupling

Speech decoupling is a common characteristic of synthetic forged speech but remains rare in speech forgery detection. [Zhu *et al.*, 2023] proposed a speech synthesis method based on multi-factor decoupling, which decomposes speech into multiple representations to obtain expressive synthesized speech. [Yang *et al.*, 2022a] utilized mutual information to decouple four representations and enhance the decoupling performance through adversarial learning. Feature decoupling is often implemented using autoencoders.There are also related studies in other fields that use decoupling mechanism to further improve model performance [Hu *et al.*, 2023; Yang *et al.*, 2022b]. [Feng *et al.*, 2018] proposed a dual-swapping technique to decouple the image representation. [Qian *et al.*, 2020] employed three encoders to respectively encode content, pitch, and rhythm information.

In this study, we employ speech feature decoupling for the task of detecting speech forgery. By combining synthetic speech samples with real speech samples, we apply fully- and self-supervised mechanisms to separate human voice features from background sound features to effectively mitigate redundant interference in speech forgery detection.

3 Methodology

As mentioned previously, human voices and background environmental sounds possess distinct characteristics. The former exhibits specific traits such as volume, rate, timbre, intonation, and tone, which are essential for identifying unique characteristics. On the other hand, background environmental sounds tend to be diverse and noisy. Contrary to existing methods that utilize mixed representations for speech forgery detection, our approach aims to amplify these distinct characteristics of human voices and background environmental sounds to detect forged speech. To obtain decoupled human voice features and background environmental sound features, we employ a swapping strategy, which involves fully supervising a synthetic speech pair and self-supervising a real speech pair, as outlined in Section 3.1. Subsequently, temporal-dimension and spectral-dimension aggregations are applied to enhance the human voice-specific features and continuity-related background features. The Heterogeneous Graph Attention Network is then employed to detect speech forgery, utilizing the enhanced features in the form of a built heterogeneous speech graph, as described in Section 3.2.

3.1 Discriminative Feature Decoupling

In this section, we employ complete supervision and self-supervision techniques to disentangle the prominent human voice feature and the background environmental sound feature in both synthetic speech and real speech, respectively.

Synthetic Speech Guided Decoupling

Different from conventional feature decoupling frameworks [Higgins *et al.*, 2017; Feng *et al.*, 2018], we propose a novel approach where two separate encoders, denoted as E_h and E_b, are designed to extract the human voice feature and the background environmental sound feature, respectively. The extracted features are then combined and used to generate synthetic speech with a single decoder D_{mix}.

To supervise the independent learning of the two encoders and the decoder in decoupling the human voice feature and the background environmental sound feature, we first synthesize several speech samples ($\widetilde{S}_{h_1b_1}$, $\widetilde{S}_{h_2b_2}$, $\widetilde{S}_{h_1b_2}$, $\widetilde{S}_{h_2b_1}$) by combining human voices (h_1 and h_2) and background environmental sounds (b_1 and b_2).

Using the synthetic speech samples, the input pairs ($\widetilde{S}_{h_1b_1}$, $\widetilde{S}_{h_2b_2}$) are fed into the encoders E_h and E_b, respectively. Through encoding, we obtain the human voice features and background sound features $[\widetilde{F}_{h_1}, \widetilde{F}_{b_1}]$ and $[\widetilde{F}_{h_2}, \widetilde{F}_{b_2}]$. As depicted in Figure 1(a), the background features are swapped, yielding new features $[\widetilde{F}_{h_1}, \widetilde{F}_{b_2}]$ and $[\widetilde{F}_{h_2}, \widetilde{F}_{b_1}]$, which are then inputted into the decoder D_{mix} to reconstruct $\widetilde{S}'_{h_1b_2}$ and $\widetilde{S}'_{h_2b_1}$. Consequently, the decoupling capability of E_h and E_b is learned by minimizing the reconstruction loss $\mathcal{L}_1$ between the decoded speeches ($\widetilde{S}'_{h_1b_2}$, $\widetilde{S}'_{h_2b_1}$) and the corresponding synthetic speeches ($\widetilde{S}_{h_1b_2}$, $\widetilde{S}_{h_2b_1}$) as follows:

$$\begin{aligned} \mathcal{L}_1 &= ||\widetilde{S}_{h_1b_2} - \widetilde{S}'_{h_1b_2}||^2 + ||\widetilde{S}_{h_2b_1} - \widetilde{S}'_{h_2b_1}||^2, \\ \widetilde{S}'_{h_1b_2} &= D_{mix}([\widetilde{F}_{h_1}, \widetilde{F}_{b_2}]), \widetilde{S}'_{h_2b_1} = D_{mix}([\widetilde{F}_{h_2}, \widetilde{F}_{b_1}]), \\ \widetilde{F}_{h_1}, \widetilde{F}_{h_2} &= E_h(\widetilde{S}_{h_1b_1}, \widetilde{S}_{h_2b_2}), \\ \widetilde{F}_{b_1}, \widetilde{F}_{b_2} &= E_b(\widetilde{S}_{h_1h_2}, \widetilde{S}_{h_2b_2}). \end{aligned} \tag{1}$$

In this way, the strongly supervised information from the synthetic speeches effectively guides the encoders E_h and E_b towards acquiring the initial feature decoupling ability.

Self-supervised Real Speech Decoupling

The encoders E_h and E_b trained using synthetic speeches exhibit poor generalization capability when applied to real speeches. To address this issue, we introduce a self-supervision technique using real speech data to enhance the decoupling ability of the encoders E_h and E_b through a dual-swapping operation, as depicted in Figure 1(b).

In the primary stage, real speech pairs ($S_{h_1b_1}$, $S_{h_2b_2}$) generate a pair of hybrid outputs ($S'_{h_1b_1}$, $S'_{h_2b_2}$) by swapping the background sound features F_{b_1} and F_{b_2} of the decoupled features $[F_{h_1}, F_{b_1}]$ and $[F_{h_2}, F_{b_2}]$. In the dual stage, the hybrid outputs ($S'_{h_1b_1}$, $S'_{h_2b_2}$) are once again fed into the same encoders E_h and E_b to obtain decoupled features $[F'_{h_1}, F'_{b_2}]$ and $[F'_{h_2}, F'_{b_1}]$. Subsequently, we swap back the background sound features F'_{b_1} and F'_{b_2} and decode the swapped features $[F'_{h_1}, F'_{b_1}]$ and $[F'_{h_2}, F'_{b_2}]$ into $S''_{h_1b_1}$ and $S''_{h_2b_2}$. As a result, the supervised loss $\mathcal{L}_2$ between the decoded speeches ($S''_{h_1b_1}$ and $S''_{h_2b_2}$) and the corresponding real speeches ($S_{h_1b_1}$ and $S_{h_2b_2}$) is calculated as follows:

$$\mathcal{L}_2 = ||S_{h_1b_1} - S''_{h_1b_1}||^2 + ||S_{h_2b_2} - S''_{h_2b_2}||^2. \tag{2}$$

The dual swap reconstruction minimization method employed in this paper offers a unique form of self-supervision. Specifically, the swapping of background sound features back and forth serves to promote the separability and modularity of the resulting human voice and background sound features. Consequently, this facilitates better decoupling ability for the two encoders, E_h and E_b, on real speech samples.

During the training stage, the Synthetic Speech Guided Decoupling and Self-supervised Real Speech Decoupling methods are integrated to train the entire framework, comprising the two encoders, E_h and E_b, and the decoder, D_{mix}.

3.2 Enhanced Features Based Forgery Detection

Utilizing the decoupled feature F_h and F_b, the Discriminative Feature Enhancement module is designed to augment the distinctive characteristics of both the human voice and background sounds. Subsequently, the Heterogeneous Graph Attention Network is employed to determine whether the input speech S is forged.

Discriminative Feature Enhancement

As previously mentioned, human voices possess distinct characteristics such as rate, timbre, intonation, and tone, which are crucial for distinguishing the unique traits of specific individuals. These voice-specific features generally remain consistent over time. Hence, we aggregate the human voice-specific feature, denoted by R_h, by calculating the maximum absolute value of F_h along the temporal dimension as follows:

$$R_h = max_t(abs(F_h)), \tag{3}$$

where $\max_t()$ represents the calculation of the maximum value along the temporal dimension.

In contrast to prominent human voices, background environmental sounds are typically diverse and noisy, lacking fixed patterns or characteristics. Consequently, we employ spectral-aggregation to combine the continuity-related background feature, represented by R_b, by calculating the maximum absolute value of F_b along the spectral dimension using the following equation:

$$R_b = max_s(abs(F_b)), \tag{4}$$

where $\max_s()$ denotes the calculation of the maximum value along the spectral dimension.

Heterogeneous Attention Based Identification

Utilizing the aggregated features R_h and R_b, the fully-connected graphs $\mathcal{G}_s$ and $\mathcal{G}_t$ are formed by connecting nodes within the node groups $\{g_s^k\}_{k=1}^K$ and $\{g_t^{k'}\}_{k'=1}^{K'}$. The nodes g_s^k and $g_t^{k'}$ are extracted from the voice-specific features R_h and enhanced background feature R_b as follows:

$$\begin{aligned} g_s^k &= R_h[k], k \in \{1, 2, ..., K\}, \\ g_t^{k'} &= R_b[k'], k' \in \{1, 2, ..., K'\}, \end{aligned} \tag{5}$$

where, K and K' denote the width of R_h and height of R_b, respectively. $R_h[k]$ and $R_b[k']$ represent the k-th vector of R_h along the width dimension and the k'-th vector of R_b along the height dimension. Please refer to Figure 1(c) for an illustrative example.

Method / Dataset	Features	FoR		ASVspoof2019LA		
		ACC	EER	ACC	Min-tDCF	EER
LCNN [Wang and Yamagishi, 2021]	cqtspec	84.32 %	14.24 %	91.22 %	0.1742	6.35 %
LCNN-Attention [Lavrentyeva *et al.*, 2017]	cqtspec	84.34 %	14.22 %	87.89 %	0.1781	6.76 %
LCNN-LSTM [Lavrentyeva *et al.*, 2019]	cqtspec	85.90 %	8.78 %	87.74 %	0.1135	6.23 %
LSTM [Zhang *et al.*, 2021a]	cqtspec	86.60 %	7.87 %	84.12 %	0.1271	7.16 %
MesoInception [Szegedy *et al.*, 2015]	logspec	80.99 %	18.91 %	79.97 %	0.2386	10.02 %
MesoNet [Afchar *et al.*, 2018]	cqtspec	81.82 %	19.83 %	83.15 %	0.2192	7.42 %
ResNet18 [Zhang *et al.*, 2021b]	cqtspec	86.77 %	9.07 %	92.45 %	0.1403	6.55 %
Transformer [Zhang *et al.*, 2021c]	cqtspec	83.22 %	16.29 %	91.04 %	0.1291	7.50 %
CRNNSpoof [Chintha *et al.*, 2020]	raw waveform	85.55 %	14.12 %	78.21 %	0.3126	15.66 %
RawNet2 [Tak *et al.*, 2021b]	raw waveform	87.37 %	7.71 %	93.13 %	0.1322	4.35 %
RawGAT-ST [Tak *et al.*, 2021a]	raw waveform	89.81 %	7.21 %	94.07 %	0.0443	1.39 %
AASIST [Jung *et al.*, 2022]	raw waveform	86.56 %	12.54 %	95.18 %	0.0347	1.13 %
DEEM (Ours)	raw waveform	**95.27 %**	**4.19 %**	**96.22 %**	**0.0287**	**1.07 %**

Table 1: Comparison experiment between the proposed DEEM and twelve existing methods on two widely-used datasets. 'ACC' and 'EER' denote the accuracy and equal error rate, respectively. Min-tDCF [Todisco *et al.*, 2019] is a metric that reflects the rate at which real speech samples are classified as forged samples. The **best performance** is indicated in bold.

Subsequently, the graphs $\mathcal{G}_s$ and $\mathcal{G}_t$ are merged to form the heterogeneous speech graph $\mathcal{G}_{st}$ by connecting all the nodes between the two graphs. Using $\mathcal{G}_{st}$ as input, a widely employed Heterogeneous Graph Attention Network is employed to extract discernible features for speech forgery detection. The well-known Cross-Entropy loss function is utilized to supervise the learning process of the Heterogeneous Graph Attention Network.

4 Experiments

In the experiment section, we begin by providing a concise overview of the dataset, parameters, and settings. Subsequently, we perform a comparative analysis between the proposed DEEM method and several state-of-the-art (SOTA) techniques. Furthermore, a comprehensive ablation study is conducted on various components to validate the efficacy of the proposed discriminative feature decoupling enhancement framework for speech forgery detection.

4.1 Datasets

Synthetic Speech Dataset. In this study, an auxiliary synthetic dataset is employed to achieve the decoupling of speech features. Specifically, we utilize publicly available datasets, namely LibriSpeech ASR [Panayotov *et al.*, 2015] and Nonspeech [Hu and Wang, 2010], to generate synthetic speech samples consisting of foreground human voices and background environmental sounds. These speech samples are accompanied by meticulous annotations, facilitating the decoupling of human voice feature and background sound feature during the training process.

- The **LibriSpeech ASR corpus** is a substantial and meticulously curated English speech dataset sourced from LibriVox audiobooks. With approximately 1000 hours of precisely segmented and aligned speech data, it serves as an optimal auxiliary dataset for the purposes of this study.
- **Nonspeech** is a collection of 100 diverse environmental audio recordings that enhance the noise set used in this study. Encompassing a wide range of real-life scenarios, it offers a rich assortment of background sound sources essential for our research.

Based on the selected foreground dataset, the LibriSpeech ASR corpus, consisting of N_h data, and the background dataset, Nonspeech, consisting of N_b data, we employ a fixed window size, denoted as w, to crop the audio data. The window's position within the audio data is determined using a random variable p, ranging from 0 to $L - w$, where L represents the length of the audio data. Additionally, we introduce another random variable v that takes values between 0 and 1 to control the intensity of the background sound during the audio stacking process. Consequently, by combining any foreground human voice h_i with a corresponding background environmental sound b_j, we generate a new mixed speech denoted as $\widetilde{S}_{h_i b_j}$.

This methodology enables the creation of $N_h \times N_b$ mixed audio samples by leveraging N_h foreground and N_b background audio samples, thereby enhancing the diversity of the dataset and fulfilling the training requirements.

Speech Forgery Benchmark Dataset. In the experimental section, this study utilizes two representative speech forgery detection datasets to evaluate the performance of the proposed algorithm.

- **FoR** [Reimao and Tzerpos, 2019] comprises an extensive collection of over $87,000$ synthetic speeches generated by advanced deep learning systems, along with over $111,000$ real speeches sourced from diverse origins. Synthetic speech that closely resembles genuine speech is generated using cutting-edge techniques such as DeepVoice3 and Google Wavenet. The dataset enhances diversity and generalization capabilities by incorporating various speech sources, speakers, recording devices, environments, and accents. In this paper, the experiments utilize the standard version of the FoR

Method	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
RawGAT-ST	1.19	0.33	0.03	1.54	0.41	1.54	0.14	0.14	1.03	0.67	1.44	3.22	0.62
AASIST	0.80	0.44	0.00	1.06	0.31	0.91	0.10	0.14	0.65	0.72	1.52	3.40	0.62
DEEM (Ours)	**0.06**	0.37	0.04	**0.39**	**0.17**	**0.33**	0.26	**0.06**	**0.24**	1.16	1.55	**2.71**	1.40

Table 2: Equal error rate (%) comparison between the DEEM method and two top SOTA methods (RawGAT-ST [Tak *et al.*, 2021a] and AASIST [Jung *et al.*, 2022]) on various unknown forgery categories in ASVspoof2019LA dataset.

dataset. It is worth noting that the standard version exhibits an uneven distribution of audio sample lengths, which presents a significant challenge.

- **ASVspoof 2019 LA** [Todisco *et al.*, 2019] serves as a dataset specifically designed for ASV anti-spoofing purposes. It encompasses both real and synthetic speech, produced using 17 TTS (Text-to-Speech) and VC (Voice Conversion) algorithms. The training and development stages involve six known attacks, while the evaluation phase incorporates 11 unknown attacks. Notably, the ASVSpoof2019LA dataset stands out for encompassing a wide array of complex spoofing algorithms in speech synthesis and voice conversion, thereby posing unique challenges for audio spoofing detection.

4.2 Parameters and Experiment Settings

The proposed DEEM model is implemented in PyTorch and evaluated on an NVIDIA Tesla V100 GPU. The decoupling module of DEEM utilizes two encoders that undergo pre-training with the swapping decoupling strategy. Subsequently, these parameters are frozen during the training of the DEEM model. To ensure consistency, all audio datasets are processed uniformly, with each sample limited to a duration of 4 seconds. Audio clips exceeding this duration are truncated, while shorter clips are padded with silence. In the decoupled training phase, a learning rate adjustment strategy is employed, with an initial learning rate set to 0.0001. If the loss value does not exhibit a significant reduction after five consecutive training iterations, the learning rate is decreased. The Adam optimizer is utilized, and the training process is carried out for 150 epochs, employing the mean squared error (MSE) loss function. In the subsequent classification training phase, the same learning rate and optimizer settings are applied. The training is performed for 160 epochs, employing the cross-entropy loss function.

The feature extractor utilized in this study is the same as the feature extractor employed in RawNet2 [Tak *et al.*, 2021b]. This feature extractor is responsible for converting the original audio into shallow features to be used in subsequent decoupling operations. The decoupling encoders encompass two identical residual sequence blocks. Each sub-encoder consists of six cascaded Residual blocks, facilitating the decoupling of the human voice and background environment sound features. The decoder involved in the decoupling training phase comprises two linear layers along with a middle deconvolution layer.

The metrics we adopted include Accuracy (ACC), Equal error rate (ERR), and Min-tDCF [Todisco *et al.*, 2019]. Min-tDCF is the minimum tandem detection cost function, which reflects the rate at which real speech samples are classified as forged samples. Due to the lack of relevant data for automatic speaker verification in the FoR dataset, it is impossible to calculate the corresponding Min-tDCF metric, Min-tDCF can only be tested on the ASVspoof2019LA dataset.

4.3 Performance Comparison with SOTA

In this section, we conduct a comparative analysis of twelve commonly used methods in the field of speech forgery detection on the FoR dataset and the ASVSpoof2019LA dataset. As illustrated in Table 1, the proposed DEEM algorithm demonstrates superior performance across all metrics.

On the FoR dataset, the DEEM algorithm achieves an Equal Error Rate (EER) index of 4.19%, which is lower than the other algorithms, and the Accuracy (ACC) index attains the highest value of 95.27%, while the ACC index of the other algorithms is all below 90%. Furthermore, on the ASVSpoof2019LA dataset, the DEEM algorithm achieves an ACC of 96.22%, an EER index of 1.07%, and a min-tDCF value of 0.0287. In comparison to other algorithms, its performance on this dataset is even more remarkable. Hence, the DEEM algorithm proposed in this paper exhibits the best overall performance and possesses evident advantages in handling the uneven time length distribution of FoR data and the presence of multiple complex spoofing types in ASVSpoof2019LA.

Additionally, we conduct more detailed performance tests on various unknown categories within the ASVSpoof2019LA dataset and compare the proposed DEEM method with the current mainstream methods displaying excellent performance. As depicted in Table 2, the DEEM approach outperforms the RawGAT-ST and AASIST methods in most categories. Notably, in categories A07, A10, A11, A12, A14, A15, and A18, the proposed method significantly surpasses the other two approaches. Moreover, in categories A08, A09, A13, and A17, the performance of the proposed method is comparable to that of the other two methods, indicating the robustness of the proposed approach in these categories. Finally, in the A16 and A19 categories, although the performance of the proposed method is slightly inferior to RawGAT-ST and AASIST, the difference is not substantial. These two categories involve waveform splicing and filtering, where the proposed DEEM algorithm may not be as effective due to its decoupling design, which primarily aims to address the fusion of foreground and background sounds. Therefore, further in-depth research is required to improve the performance in these areas.

After analyzing on different forgery categories of ASVSpoof2019LA dataset, it becomes evident that the proposed DEEM method effectively detects vocoder and GAN spoofing types, along with their diverse variations.

Operation	FoR		ASVspoof2019LA		
	ACC	EER	ACC	EER	Min-tDCF
$A_s(F_h)$ +$A_s(F_b)$	80.71%	9.34%	89.39%	6.61%	0.1832
$A_t(F_h)$ +$A_t(F_b)$	77.94%	18.98%	78.15%	11.08%	0.3706
$A_t(F_h)$ +$A_s(F_h)$	90.25%	7.34%	94.39%	2.60%	0.0786
$A_t(F_b)$ +$A_s(F_b)$	85.32%	8.94%	92.71%	4.88%	0.1806
$A_t(F_b)$ +$A_s(F_h)$	77.45%	16.22%	92.58%	5.68%	0.1765
$A_t(F_h)$ +$A_s(F_b)$	**95.27%**	**4.19%**	**96.22%**	**1.07%**	**0.0287**

Table 3: Results of the ablation study on different feature aggregation strategies. Aggregating features of human voice feature F_h and background feature F_b along the temporal-dimension $A_t()$ and spectral-dimension $A_s()$.

The remarkable performance of the DEEM in forged speech detection tasks can be attributed to its design of deep feature decoupling. By separating the foreground and background sounds and eliminating redundant interference, DEEM can accurately capture and aggregate discriminative information for forged speech detection tasks, thus substantially enhancing the detection accuracy.

4.4 Ablation Study

In order to validate the efficacy of our significant contribution, namely the decoupled and enhanced features, several ablation studies were performed. These studies investigate various feature aggregation strategies and different back-end classifiers.

Feature Aggregation Strategy

For the decoupled human voice feature F_h and background environmental sound feature F_b, we employ temporal- and spectral-dimension aggregation strategies to extract human voice-specific features and background continuity-related features from the background sound, respectively.

To verify the effectiveness and rationality of our aggregation strategy, we compare different combinations of aggregating features F_h and F_b along various dimensions. The results obtained with different aggregation strategies for F_h and F_b are given in Table 3. In this context, $A_t()$ and $A_s()$ represent the aggregation feature along the temporal- and spectral-dimensions, respectively.

From the results presented in Table 3, we observe that the combination $A_t(F_h) + A_s(F_b)$ achieves the best performance, which confirms the effectiveness of our proposed feature enhancement strategy. Additionally, we note that $A_t(F_h)+A_s(F_h)$ achieves the second-best performance, suggesting that prominent human voice plays a vital role in speech forgery detection.

Another significant finding is that the aggregation operation $A_t(F_b) + A_s(F_h)$ yields the poorest performance, as it employs the opposite aggregation operation. This finding further confirms that aggregating voice-specific features in the temporal dimension, coupled with aggregating continuity-related background sound features in the spectral dimension, represents the optimal operation for speech forgery detection.

Effectiveness of Decoupled and Enhanced Features

To assess the efficacy of the decoupled and enhanced features, we employed them as input features for several existing methods, namely LCNN [Wang and Yamagishi, 2021], RawNet2 [Tak *et al.*, 2021b], RawGAT-ST [Tak *et al.*, 2021a], and AASIST [Jung *et al.*, 2022]. The performance differences between these methods using mixed deep features and the decoupled and enhanced features are summarized in Table 4.

Method	FoR		ASVspoof2019LA		
	ACC	EER	ACC	EER	Min-tDCF
LCNN	+10.45%	-10.78%	+4.21%	-0.25%	-0.0284
RawNet2	+4.32%	-4.29%	+2.36%	-0.46%	-0.0008
RawGAT-ST	+4.56%	-3.04%	+1.27%	-0.04%	-0.0026
AASIST	+8.71%	-8.35%	+1.04%	-0.07%	-0.0062

Table 4: The impact of employing our decoupled and enhanced features to replace the original mixed embedding in existing methods.

It can be observed that LCNN exhibited the most substantial improvement on the FoR dataset, with an increase in accuracy (ACC) by 10.45% and a decrease in equal error rate (EER) by 10.78%. Other methods also demonstrated some level of improvement. Likewise, in the evaluation of the ASVSpoof2019LA dataset, the employment of deep decoupling features resulted in enhanced EER and minimum tandem detection cost function (Min-tDCF) performance across the methods. LCNN achieved a reduction in the EER by 0.25%, while the Min-tDCF reduced by 0.0284%. RawGAT-ST and AASIST also exhibited a certain degree of reduction in EER and Min-tDCF indicators. These results suggest that the decoupled and enhanced feature exhibits broad generalization capabilities across different classifiers.

5 Conclusion

In this paper, we propose a Discriminative fEature dEcoupling enhanceMent framework (DEEM) for speech forgery detection. DEEM effectively separates speech into two distinct feature maps: the human voice feature map and the background sound feature map, using a swapping decoupling strategy. By applying temporal dimension aggregation on the human voice feature map and spectral-dimension aggregation on the background sound feature map, we enhance the voice-specific features and continuity-related background features, respectively. These enhanced features are integrated into a fully-connected heterogeneous graph, and a heterogeneous graph attention network is employed to extract forgery features for speech detection. Experimental results demonstrate that DEEM achieves significant accuracy improvement. In future research, we will prioritize the investigation of diverse distinctive features for multimedia forgery detection.

Acknowledgments

This work is funded by National Key Research and Development Project (Grant No: 2022YFB2703100), Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" (No. 2022R52046).

References

[Afchar *et al.*, 2018] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[Agarwal and Verma, 2020] Ritu Agarwal and Om Prakash Verma. An efficient copy move forgery detection using deep learning feature extraction and matching algorithm. *Multimedia Tools and Applications*, 79(11-12):7355–7376, 2020.

[Aihara *et al.*, 2013] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8037–8040. IEEE, 2013.

[Alzantot *et al.*, 2019] Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. Deep residual neural networks for audio spoofing detection. *arXiv preprint arXiv:1907.00501*, 2019.

[Arık *et al.*, 2017] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, pages 195–204. PMLR, 2017.

[Bigorgne *et al.*, 1993] D Bigorgne, Olivier Boeffard, B Cherbonnel, Françoise Emerard, Danielle Larreur, JL Le Saint-Milon, I Metayer, Christel Sorin, and S White. Multilingual psola text-to-speech system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 187–190. IEEE, 1993.

[Chen *et al.*, 2024] Jiawei Chen, Lin Chen, Jiang Yang, Tianqi Shi, Lechao Cheng, Zunlei Feng, and Mingli Song. Life regression based patch slimming for vision transformers. *Neural Networks*, page 106340, 2024.

[Chintha *et al.*, 2020] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1024–1037, 2020.

[Desai *et al.*, 2009] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3893–3896. IEEE, 2009.

[Feng *et al.*, 2018] Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, and Mingli Song. Dual swap disentangling. In *Advances in Neural Information Processing Systems*, pages 5894–5904, 2018.

[Guillaro *et al.*, 2023] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20606–20615, 2023.

[Helander *et al.*, 2008] Elina Helander, Jan Schwarz, Jani Nurminen, Hanna Silen, and Moncef Gabbouj. On the impact of alignment on voice conversion performance. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[Hu and Wang, 2010] Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

[Hu *et al.*, 2023] Kaiwen Hu, Jing Gao, Fangyuan Mao, Xinhui Song, Lechao Cheng, Zunlei Feng, and Mingli Song. Disassembling convolutional segmentation network. *International Journal of Computer Vision*, 131(7):1741–1760, 2023.

[Jung *et al.*, 2022] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

[Kaneko and Kameoka, 2018] Takuhiro Kaneko and Hirokazu Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE, 2018.

[Koptyra and Ogiela, 2020] Katarzyna Koptyra and Marek R Ogiela. Imagechain—application of blockchain technology for images. *Sensors*, 21(1):82, 2020.

[Lai *et al.*, 2019] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. Assert: Anti-spoofing with squeeze-excitation and residual networks. *arXiv preprint arXiv:1904.01120*, 2019.

[Lavrentyeva *et al.*, 2017] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Interspeech*, pages 82–86, 2017.

[Lavrentyeva *et al.*, 2019] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*, 2019.

[Li *et al.*, 2021] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Replay and synthetic speech detection with res2net architecture. In *ICASSP*

2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 6354–6358. IEEE, 2021.

[Morise *et al.*, 2016] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[Oord *et al.*, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Panayotov *et al.*, 2015] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[Qian *et al.*, 2020] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020.

[Reimao and Tzerpos, 2019] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10. IEEE, 2019.

[Sun *et al.*, 2016] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Tak *et al.*, 2021a] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654*, 2021.

[Tak *et al.*, 2021b] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

[Todisco *et al.*, 2019] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[Wang and Yamagishi, 2021] Xin Wang and Junich Yamagishi. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326*, 2021.

[Wang *et al.*, 2017] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[Wu *et al.*, 2018] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[Yang *et al.*, 2022a] SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *arXiv preprint arXiv:2208.08757*, 2022.

[Yang *et al.*, 2022b] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022.

[Yoshimura, 2002] Takayoshi Yoshimura. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems. *PhD diss, Nagoya Institute of Technology*, 2002.

[Zhang *et al.*, 2021a] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. An initial investigation for detecting partially spoofed audio. *arXiv preprint arXiv:2104.02518*, 2021.

[Zhang *et al.*, 2021b] You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021.

[Zhang *et al.*, 2021c] Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. Fake speech detection using residual network with transformer encoder. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, pages 13–22, 2021.

[Zheng *et al.*, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021.

[Zhu *et al.*, 2023] Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, and Lei Xie. Multi-speaker expressive speech synthesis via multiple factors decoupling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.