Multi-teacher Knowledge Distillation for Spoof Speech Detection

Zirui Chen, Rui Wang, Tianxiang Luo, Yeling Tang, Bo Wang*

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116081, Dalian, China

ABSTRACT

The escalating issue of telecommunication fraud and the spread of disinformation, instigated by spoof speech, urgently calls for efficient and broadly applicable detection methods. However, existing detection methods face significant challenges: they struggle to effectively accommodate the distinct characteristics of synthesized speech (TTS) and converted speech (VC[1]), while also suffering from the limitations of relying on emotion features for VC detection. To address these dual challenges, this paper proposes a detection method based on multi-teacher knowledge distillation. Our approach involves separately training dedicated TTS and VC teacher models. These teachers are designed to capture specific forgery cues: the TTS teacher extracts emotion-deficient features inherent in synthesized speech, and the VC teacher identifies spectral distortion features characteristic of converted speech. The core distillation process then transfers this specialized knowledge into a single, lightweight student model using mean square error loss. Experimental validation demonstrates the efficacy of our method: the distilled student model achieves remarkably low Equal Error Rates (EER) of 0.38% for synthesized speech detection and 4.60% for converted speech detection. This represents a significant advancement, surpassing the baseline model's performance and achieving an overall EER of 1.84% on the ASVspoof 2019 LA test set.[2] Furthermore, ablation experiments and a generalizability analysis substantiate the framework's robustness, offering fresh perspectives for deep fake detection across various forgery types.

Keywords: Audio deepfake detection, Multi-teacher knowledge distillation, OC-Softmax

1. INTRODUCTION

The rapid advancement of deepfake speech technologies has led to the creation of highly realistic spoof speech. This poses significant threats to privacy and security, which highlights the urgency of research on deepfake speech detection. The goal of deepfake speech detection is to distinguish between genuine and manipulated speech by identifying forgery traces through machine or deep learning algorithms, ensuring information security and maintaining social stability.

Early studies relied on handcrafted short-term spectral features and long-term spectral features. These leveraged local spectral characteristics and long-term dependencies for discrimination. Baseline models in ASVspoof challenges validated the effectiveness of such features, but their representational capacity was limited by manual priors. The integration of prosodic and emotional features[3] further enhanced cross-dataset generalization. Classifier architectures evolved from traditional GMM/SVM to deep learning frameworks, with lightweight CNNs (LCNN), ResNets, and graph attention networks (e.g. AASIST[4]) becoming mainstream. The AASIST model achieved state-of-the-art performance in ASVspoof 2021 through heterogeneous spatio-temporal graph attention mechanisms. Its end-to-end architecture jointly optimized feature extraction and classification modules, significantly improving detection efficiency.

Despite notable progress, critical challenges persist: 1) Due to the distinct generation mechanisms between text-to-speech (TTS) and voice conversion (VC) systems, single detection models often struggle to accurately

Further author information: (Send correspondence to Bo Wang)
Bo Wang: E-mail: bowang@dlut.edu.cn
Zirui Chen: E-mail: chenzirui@mail.dlut.edu.cn

International Conference on Advances in Computer Science, Information Technology, and Communications (CSITC 2025), edited by Yong Yue, Iickho Song, Jixin Ma, Proc. of SPIE Vol. 13969, 1396904
 · 0277-786X · doi: 10.1117/12.3093006

capture discriminative features applicable to both synthetic and converted speech, leading to suboptimal performance in detecting at least one category of spoof audio. 2) Emotion-based detection methods exhibit inherent limitations. While TTS often fails to incorporate natural emotional cues, the emotional characteristics of source speech in VC are preserved and transferred to target utterances during the conversion process. This discrepancy results in compromised generalization capacity of emotion-driven spoofing detection frameworks.

To address these issues, this paper proposes a multi-teacher knowledge distillation framework for spoof speech detection. The proposed method simultaneously enhances detection performance across both TTS and VC domains while overcoming the insufficiency of single emotional feature representations. This provides a more comprehensive and generalized solution for speech deepfake detection.

2. METHODOLOGY

This section outlines our proposed spoof speech detection method based on multi-teacher knowledge distillation and its core modules. The overall framework of the forged speech detection method, based on multi-teacher knowledge distillation, is depicted in Fig 1. It comprises three parts: the student model, the TTS teacher model, and the VC teacher mode.

In this method, we opt to combine a simple traditional handcrafted feature, the Linear Frequency Cepstrum Coefficient (LFCC), with a ResNet18 classification network to form the student model. Concurrently, we introduce a One-Class Softmax (OC-Softmax)[5] loss function after the fully connected layer of the student model. This enhances the model's generalization ability, enabling it to handle different types of forgery algorithms as well as unknown forgery algorithms.

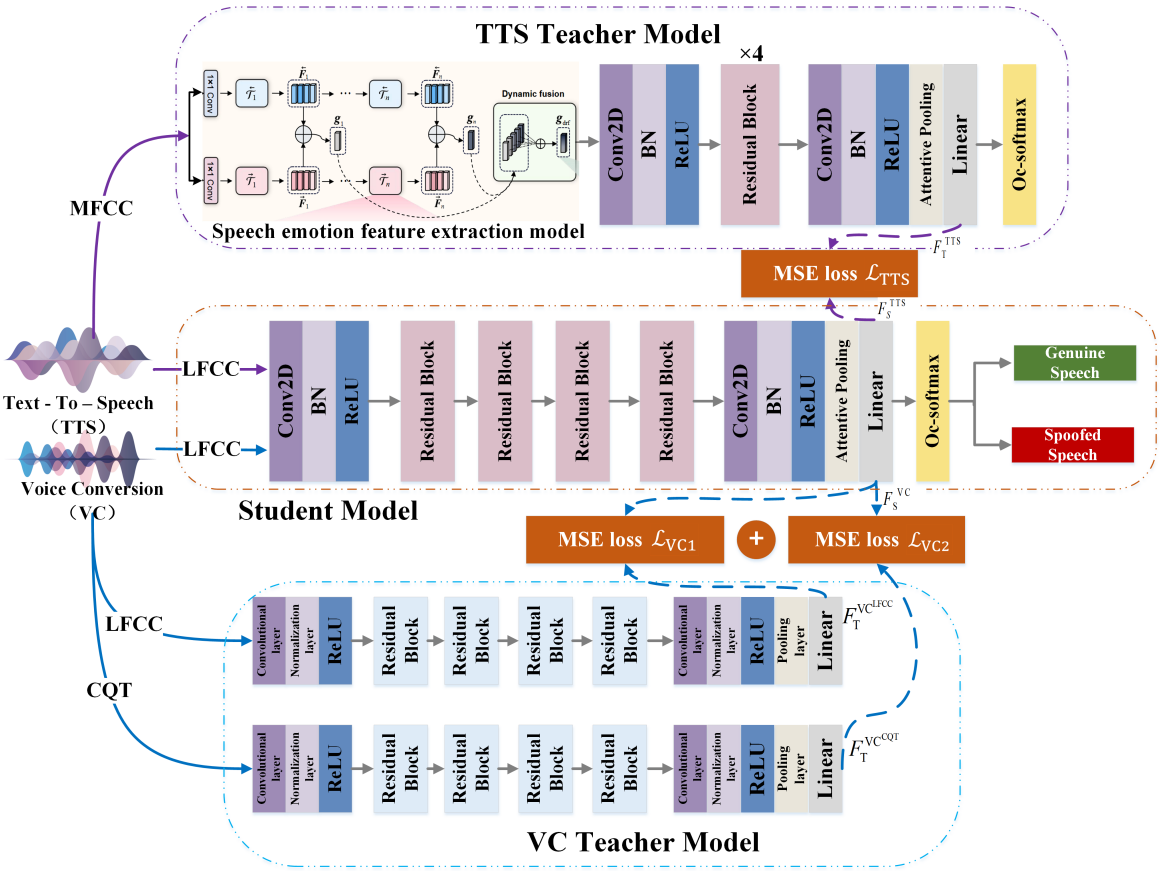

Figure 1: The overall framework of multi-teacher knowledge distillation for audio deepfake detection

2.1 TTS Knowledge Distillation Module

The TTS teacher model is trained using synthetic and real speech from the ASVspoof 2019 LA training set, enabling it to learn the unique artifact features in synthetic speech. The role of the TTS knowledge distillation module is to use the TTS teacher model to impart the unique artifact information of synthesized speech to the student model, thereby enhancing the student model's detection performance for synthesized speech. The input for TTS speech synthesis technology[6] is purely textual information, and the emotional expression of such information can vary in different contexts. Moreover, the expression of emotion often relies on multimodal information such as facial expressions and body language. Given that speech synthesis algorithms rely solely on textual inputs, which are challenging to capture, it becomes difficult for speech synthesis techniques to correctly generate natural emotional information. In this paper, we found through playback tests that the synthesized speech in publicly available datasets in the field of fake speech detection tends to sound neutral and lacks emotional fluctuations. Therefore, the TTS teacher model in this method adopts emotional features as the key features for distinguishing between authentic and fake speech.

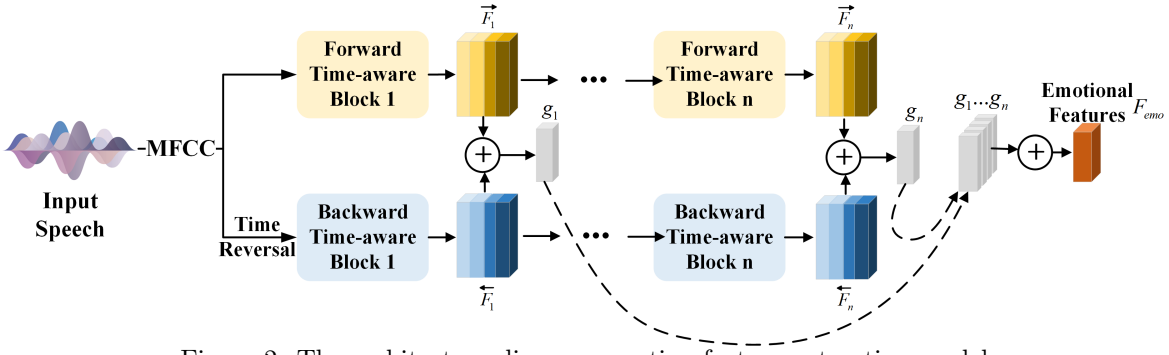

Figure 2: The architecture diagram emotion feature extraction model

The TTS teacher model employs a bidirectional time-aware multi-scale emotion recognition model (TIM-Net)[7] as the speech emotion feature extractor, as depicted in Figure 2. Given that the assessment of speech emotion requires the integration of global contextual information, the TIM-Net model is designed with a bidirectional architecture. This architecture processes the forward and backward signals of the input speech through two branches, merging the information from the preamble and postamble of the speech, respectively. Initially, the TIM-Net model extracts the Mel Frequency Cepstrum Coefficients (MFCC) of the speech, followed by a time inversion of the MFCC features. Subsequently, the TIM-Net model inputs the forward-ordered MFCC features and the reverse-ordered MFCC features into a forward branch and a reverse branch, each composed of multiple time-aware blocks. The structure of the time-aware blocks utilizes dilation causal convolution. Dilation convolution expands the perceptual field of the convolution, while causal convolution ensures that the convolution only accesses present and past data, preventing future data from interfering with the current prediction. The model combines the output features of each corresponding forward and reverse time-aware block to fuse the bidirectional semantic dependencies of speech, thereby enriching the contextual representation of the features. This process is formulated as follows:

$$g_i = \mathcal{G}\left(\vec{\boldsymbol{F}}_{\boldsymbol{i}} + \overleftarrow{\boldsymbol{F}}_{\boldsymbol{i}}\right) \tag{1}$$

Among them, i denotes the serial number of the time-aware block, represents the global time pooling operation averaged over the time dimension, $\vec{\boldsymbol{F}}$ signifies the output features of the forward time-aware block, and $\overleftarrow{\boldsymbol{F}}$ indicates the output features of the reverse time-aware block. Ultimately, the TIM-Net model weights and sums g_i to g_{n} to fuse multiple time-scale features, thereby better adapting to changes in speech emotion. The formula for combining g_i to g_{n} is as follows:

$$F_{emo} = \sum_{i=1}^{n} w_i \cdot g_i \tag{2}$$

Among them, n denotes the number of time-aware blocks, and w_i is a learnable parameter representing the dynamic weight of g_i. The TIM-Net model is trained using the English dataset SAVEE[8] , which encompasses seven emotions (anger, disgust, fear, happiness, neutrality, sadness, and surprise). This model serves as an emotion feature extractor for this method, applicable to both real and synthesized speech. The proposed method incorporates a ResNet18 classification network following the emotional feature extractor, where the ResNet18 is designed to learn more discriminative deep-level features embedded within the emotional characteristics. The TTS knowledge distillation module transfers synthetic speech detection knowledge from the aforementioned TTS teacher model to the student model. This knowledge transfer is implemented via the mean squared error (MSE) loss between their feature representations, formulated as:

$$\mathcal{L}_{\mathrm{TTS}} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_{\mathrm{T},i}^{\mathrm{TTS}} - F_{\mathrm{S},i}^{\mathrm{TTS}}\right\|^2 \tag{3}$$

Among them, N denotes the number of speech samples input to the TTS teacher model for TTS knowledge distillation, where $F_{T,i}^{\mathrm{TTS}}$ and $F_{\mathrm{S},i}^{\mathrm{TTS}}$ represent the feature embeddings generated by feeding the ith speech sample into the TTS teacher and the student model, respectively.

2.2 VC Knowledge Distillation Module

The VC knowledge distillation module[9] aims to teach the student model the unique artifact information of converted speech using the VC teacher model, thereby enhancing the student model's detection performance for converted speech. The VC speech conversion technique is designed to convert the source speaker's speech features into the target speaker's speech features. Its input signal is the source speaker's speech data, which contains rich emotional information about the speaker. Speech conversion technology can directly extract and retain this emotional information from the source speech, resulting in the target speech obtained by conversion being rich in emotional information. Therefore, emotional features cannot be used as the key features for converted speech detection. Meanwhile, due to the complexity of speech conversion technology, converted speech retains the semantic content, emotional information, and acoustic features of the source speech. This makes it less distinguishable from real speech.

For this reason, this method selects a different detection model, the Dual-Branch Network[10] ,as the VC teacher model for converted speech. The model, shown in Fig 3, adopts a dual-branch network structure to enhance the detection performance of forged speech through the complementarity of different features. It first extracts the LFCC and CQT of the input speech as inputs to the two branches. The LFCC, extracted based on the Short-Time Fourier Transform, is suitable for capturing transient anomalies of speech signals and is more robust to noise and channel distortion. It reflects the vocal tract characteristics of speech through cepstrum analysis and is suitable for detecting vocal tract anomalies in falsified speech. On the other hand, CQT has variable time-frequency resolution, with high frequency resolution in the low-frequency region, which is suitable for analyzing fundamental and harmonics, and high time resolution in the high-frequency region, which is suitable for capturing rapidly changing signals.

Sensitive to the fundamental and harmonic structures, the CQT can directly reflect spectral features, making it suitable for detecting spectral distortion and traces of artificial processing in forged speech. The LFCC and CQT complement each other across multiple dimensions, comprehensively capturing the forged features of the speech and improving the performance of forgery detection.

The classification networks in both branches of the Dual-Branch Network employ residual blocks from ResNet18, augmented with Convolutional Block Attention Modules to strengthen the feature representation capability of the model.

The Dual-Branch Network model incorporates multi-task learning during the training process, which includes a forgery speech detection task and a forgery type classification task. These two tasks are trained concurrently. The forgery type classification task assists the model in learning common forgery features from different types of forged speech, thereby enhancing the model's resilience against unknown forgery attacks. The model employs

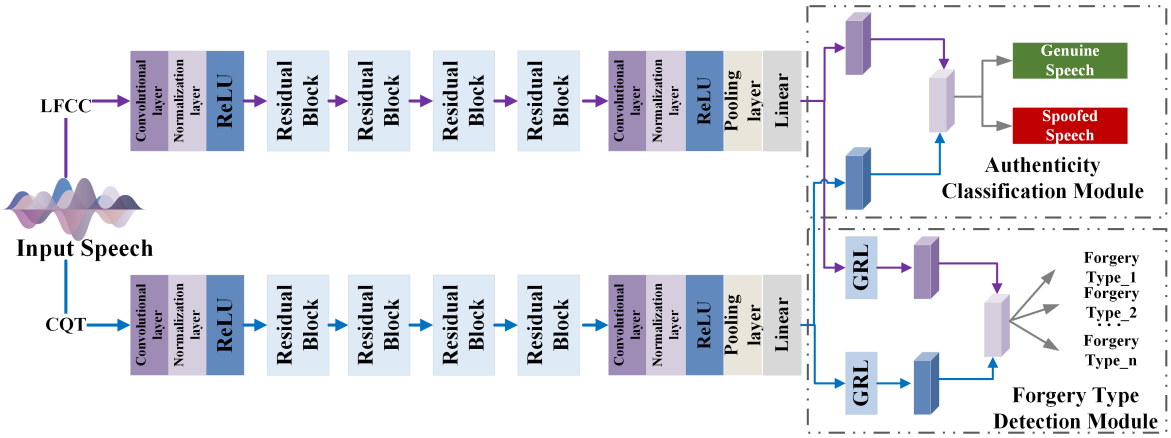

Figure 3: The structure diagram of VC teacher model

the Cross Entropy Loss function in both the authenticity classification module and the forgery type detection module. The formula for this function is as follows:

$$\mathcal{L} = -\frac{1}{N}\left(\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} * lnP_{i,k}^{LFCC} + \sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} * lnP_{i,k}^{CQT}\right) \tag{4}$$

In this formula, N denotes the number of speech samples, K represents the number of categories of speech samples, $P_{i,k}^{LFCC}$ and $P_{i,k}^{CQT}$ denote the probabilities that the $LFCC$ branch and CQT branch predict that the ith sample belongs to the kth category, respectively. $y_{i,k}$ denotes the sample label, and $y_{i,k}$ equals 1 when the ith sample belongs to the kth label.

To enable the VC teacher model to learn the unique artifact information of converted speech, this method selects only the converted type of speech as the fake data to train the Dual-Branch Network model. Due to the scarcity of converted speech in the ASVspoof 2019 LA training set, this method uniquely includes converted speech from the ASVspoof 2015 training set when training the VC teacher model, thereby enhancing the diversity of the training data.

The VC knowledge distillation module imparts the knowledge of transformed speech detection, learned in the two branches of the VC teacher model, to the student model through distillation loss. This is defined by the following loss function:

$$\mathcal{L}_{\text{vc}}^{\text{lfcc}} = \frac{1}{N}\sum_{i=1}^{N}\left\| F_{\text{T},i}^{\text{VC}^{lfcc}} - F_{\text{S},i}^{\text{VC}} \right\|^2 \tag{5}$$

$$\mathcal{L}_{\text{vc}}^{\text{cqt}} = \frac{1}{N}\sum_{i=1}^{N}\left\| F_{\text{T},i}^{\text{VC}^{cqt}} - F_{\text{S},i}^{\text{VC}} \right\|^2 \tag{6}$$

$$\mathcal{L}_{\text{VC}} = \lambda\mathcal{L}_{\text{vc}}^{\text{lfcc}} + \gamma\mathcal{L}_{\text{vc}}^{\text{cqt}} \tag{7}$$

Among them, $\mathcal{L}_{\text{vc}}^{\text{dfcc}}$ and $\mathcal{L}_{\text{vc}}^{\text{cqt}}$ denote the distillation losses between the $LFCC$ and CQT branches of the VC teacher model and the student model, respectively. N represents the number of speech samples input to the VC teacher model for VC knowledge distillation.$F_{\text{S},i}^{\text{VC}}$, $F_{\text{T},i}^{\text{VC}^{lfcc}}$ and $F_{\text{T},i}^{\text{VC}^{cqt}}$ correspond to the feature representations generated by feeding the ith speech sample into the student model, the $LFCC$ branch of the VC teacher model, and the CQT branch of the VC teacher model, respectively. The total loss of the VC knowledge distillation module is formulated as a weighted sum of $\mathcal{L}_{\text{vc}}^{\text{dfcc}}and\mathcal{L}_{\text{vc}}^{\text{cqt}}$, where λ and γ denote the weights assigned to the distillation losses of the two branches.

2.3 Overall loss function

The overall loss of the forgery speech detection method, based on multi-teacher knowledge distillation, comprises three components: the loss from TTS knowledge distillation, the loss from VC knowledge distillation, and the categorization loss of the student model. These are calculated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{\text{TTS}} + \beta\mathcal{L}_{\text{VC}} \tag{8}$$

In this formula,$\mathcal{L}_{CE}$ represents the categorization loss of the student model, while α and β are weight hyperparameters used to balance the TTS knowledge distillation loss and the VC knowledge distillation loss, respectively.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental Setup

The learning rate of this method is set to 0.0003, and the decay is set by a factor of 0.5 for each of the 10 rounds. The batch size is set to 64, and each model is trained for 100 rounds. In the VC knowledge distillation module, to equalize the contribution of the two branches in knowledge transfer, the values of the parameters λ and γ are both set to 0.5. In the overall loss function, this method sets the value of the loss weight parameter α for the TTS knowledge distillation to 1, and that of the loss weight parameter β for the VC knowledge distillation to 5.

3.2 Ablation Experiment

This method quantitatively analyzes the impact of the OC-Softmax loss function, the TTS knowledge distillation module, and the VC knowledge distillation module on the overall model performance through ablation experiments. These experiments are conducted on a subset of different forgery types from the ASVspoof 2019 LA test set, with the results shown in Table 1. Among them, the Student model refers to the LFCC_ResNet18_OC - Softmax model. Due to the simpler speech features and the shallow classification network of the LFCC_ResNet18 model, the model exhibits poor detection performance on the overall test set, with the detection EER on converted speech reaching as high as 9.57%. However, with the addition of the OC-Softmax loss function, the student model's detection performance on each class of falsified speech improves. This suggests that the classification decision boundary of the OC-Softmax loss function effectively distinguishes real speech from various types of forged speech, thereby enhancing the model's ability to generalize to different types of forgery attacks. After adding only the TTS knowledge distillation module, the student model's forgery detection EER for synthetic speech decreases from 0.90% to 0.67%. This indicates that the TTS knowledge distillation module successfully transfers the TTS teacher model's knowledge about synthetic speech detection to the student model. Similarly, the student model's EER for forgery detection of transformed speech decreases from 6.72% to 5.34% when only the VC knowledge distillation module is added, demonstrating that the student model has learned the VC teacher model's knowledge about the detection of transformed speech. When both the TTS and VC knowledge distillation modules are added, the student model's detection performance for both synthesized and converted speech improves further compared to when either knowledge distillation module is added alone. The detection EERs for synthesized speech, converted speech, and the overall test set are 0.38%, 4.60%, and 1.84%, respectively.

Table 1: The ablation experiment results (EER) of the multi-teacher knowledge distillation model on the ASVspoof 2019 LA test set

Test	Model				
Set	LFCC_ResNet18	Student	Student+KD$_{TTS}$	Student+KD$_{VC}$	Student+KD$_{TTS+VC}$
TTS	2.97%	0.90%	0.67%	1.45%	0.38%
VC	9.57%	6.72%	9.84%	5.34%	4.60%
ALL	4.70%	2.87%	2.80%	2.30%	1.84%

The results of the ablation experiments demonstrate that each module selected for this method positively contributes to the overall model's forgery detection capability.

3.3 TTS & VC Knowledge Distillation Module Generalizability Analysis Experiments

To verify the generality of the TTS and VC knowledge distillation modules of this method for different student models, we replaced the aforementioned student model with four other types of falsified speech fabrication detection models: XLS-R+SpecRNet, XLS-R+Resnet18, the AASIST end-to-end model, and Dual-Branch Network. This was done to test the performance enhancement effect of this method's knowledge distillation module on different types of models for fake speech detection, as the experimental results are shown in Table 2. These four pseudo-speech forgery detection models already exhibit excellent forgery detection performance on their own, and their performance is further improved after adding the TTS and VC knowledge distillation modules of this method.These experimental results demonstrate that the TTS and VC knowledge distillation can be viewed as a framework that can be applied to various fake speech detection models to simultaneously improve the fake detection performance of the original models for both synthesized speech and converted speech.

Table 2: The experimental results (EER) of the generalization analysis of the TTS&VC knowledge distillation modules

Model	Test Data					
	19 LA TTS	△	19 LA VC	△	19 LA ALL	△
AASIST	0.69%		2.75%		1.39%	
AASIST+ $\mathbf{KD_{TTS-VC}}$	**0.57**%	↓ 0.12%	**2.09**%	↓ 0.66%	**1.20**%	↓ 0.19%
Dual-Branch	0.75%		1.59%		0.87%	
Dual-Branch+ $\mathbf{KD_{TTS-VC}}$	**0.21**%	↓ 0.54%	**1.38**%	↓ 0.21%	**0.59**%	↓ 0.28%
XLS-R+Resnet18	0.39%		0.26%		0.37%	
XLS-R+SpecRNet+ $\mathbf{KD_{TTS-VC}}$	**0.31**%	↓ 0.08%	**0.20**%	↓ 0.06%	**0.26**%	↓ 0.11%
XLS-R+SpecRNet	0.29%		0.27%		0.29%	
XLS-R+SpecRNet+ $\mathbf{KD_{TTS-VC}}$	**0.11**%	↓ 0.18%	**0.24**%	↓ 0.03%	**0.16**%	↓ 0.13%

3.4 Comparative Experiment

This method uses the relatively simple student model LFCC_ResNet18_OC-Softmax+KD (denoted as Student_1+KD), trained using TTS and VC knowledge distillation, and the most effective student model XLS-R+SpecRNet+KD (denoted as Student_2+KD). These are compared with other existing forgery detection models that also focus on the detection performance of each forgery algorithm and the results are shown in Table 3. The two student models of this method outperform the detection models of the same type. In particular, the Student_2+KD model shows optimal detection performance on most of the forgery algorithms, which fully reflects the superiority of the TTS and VC knowledge distillation of this method.

Table 3: The comparative experiment results (EER%) between the multi-teacher knowledge distillation model and existing spoof detection models on the ASVspoof 2019 LA dataset

Model	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	ALL
RawNet2[11]	3.93	4.50	0.15	2.97	1.65	3.78	0.24	0.90	3.10	1.04	6.74	14.98	1.85	4.62
IAIF-KNN[12]	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.50	4.56	4.46	3.50	3.50
Student_1+KD	0.14	0.10	0.10	0.79	**0.10**	0.34	0.26	0.58	0.73	0.38	9.23	1.28	0.66	1.84
Wav2+AASIST[13]	0.11	0.53	0.02	0.42	0.16	0.14	**0.02**	**0.04**	0.11	0.19	3.24	2.79	3.15	0.90
TC[14]	0.23	0.69	0.23	0.67	0.23	0.23	0.29	0.23	0.24	0.59	1.10	2.00	1.10	0.84
TSF[15]	0.12	0.26	0.12	**0.26**	0.15	0.12	0.15	0.37	0.49	0.19	2.64	0.67	0.41	0.77
DFSincNet[16]	0.37	0.08	0.04	0.76	0.49	0.65	0.04	0.18	0.39	0.29	1.02	0.63	0.65	0.52
Student_2+KD	**0.04**	**0.08**	**0.02**	0.29	0.22	**0.06**	0.11	0.10	**0.06**	**0.06**	**0.24**	**0.26**	**0.34**	**0.16**

4. CONCLUSION

In this paper, we propose a forgery speech detection method based on multi-teacher knowledge distillation. This method addresses the issues of insufficient generalization ability and limitations of affective feature forgery

detection when the model simultaneously detects two types of forgery methods: speech synthesis and speech conversion. The method employs two teacher models, each focusing on the detection tasks of synthesized and converted speech, respectively. These models pass the proprietary discriminative knowledge of synthesized and converted speech to the student models through a multi-teacher knowledge distillation mechanism. A series of experimental results demonstrate that the TTS and VC knowledge distillation modules of this method can simultaneously improve the detection performance of synthetic and converted speech for different types of student models. This reflects the effectiveness and versatility of multi-teacher knowledge distillation in this method.

REFERENCES

[1] Sisman, B., Yamagishi, J., King, S., and Li, H., "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 132–157 (2020).

[2] Nautsch, A., Wang, X., Evans, N., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K. A., "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3(2), 252–265 (2021).

[3] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M. C., and Tubaro, S., "Deepfake speech detection through emotion recognition: a semantic approach," in [*ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*], 8962–8966, IEEE (2022).

[4] Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J. S., Lee, B.-J., Yu, H.-J., and Evans, N., "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)*], 6367–6371, IEEE (2022).

[5] Zhang, Y., Jiang, F., and Duan, Z., "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters* 28, 937–941 (2021).

[6] Kong, J., Kim, J., and Bae, J., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems* 33, 17022–17033 (2020).

[7] Ye, J., Wen, X.-C., Wei, Y., Xu, Y., Liu, K., and Shan, H., "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in [*ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*], 1–5, IEEE (2023).

[8] Jackson, P. and Haq, S., "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK* (2014).

[9] Sisman, B., Yamagishi, J., King, S., and Li, H., "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 132–157 (2020).

[10] Ma, K., Feng, Y., Chen, B., and Zhao, G., "End-to-end dual-branch network towards synthetic speech detection," *IEEE Signal Processing Letters* 30, 359–363 (2023).

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

[12] Muttathu Sivasankara Pillai, A. S., L. De Leon, P., and Roedig, U., "Detection of voice conversion spoofing attacks using voiced speech," in [*Nordic Conference on Secure IT Systems*], 159–175, Springer (2022).

[13] Tak, H., Todisco, M., Wang, X., Jung, J.-w., Yamagishi, J., and Evans, N., "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233* (2022).

[14] Zhang, Y., Li, Z., Lu, J., Wang, W., and Zhang, P., "Synthetic speech detection based on the temporal consistency of speaker features," *IEEE Signal Processing Letters* (2024).

[15] Wen, P., Hu, K., Yue, W., Zhang, S., Zhou, W., and Wang, Z., "Robust audio anti-spoofing with fusion-reconstruction learning on multi-order spectrograms," *arXiv preprint arXiv:2308.09302* (2023).

[16] Huang, B., Cui, S., Huang, J., and Kang, X., "Discriminative frequency information learning for end-to-end speech anti-spoofing," *IEEE Signal Processing Letters* 30, 185–189 (2023).