TJU SYSTEM DESCRIPTION TO SLT 2024 SOURCE SPEAKER TRACING CHALLENGE

Xinlei Ma[1]*, Wenhuan Lu*[1]*, Ruiteng Zhang*[1]*, Junhai Xu*[1]*, Xugang Lu*[2]*, Jianguo Wei*[1,†]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National Institute of Information and Communications Technology, Kyoto, Japan

jianguo@tju.edu.cn

ABSTRACT

Automatic speaker verification (ASV) systems face significant challenges when exposed to spoofing attacks, necessitating robust countermeasures. In this work, we focus on the source speaker verification (SSV) task, which aims to identify the source speaker hidden in spoofed speech generated by voice conversion (VC) techniques. We propose a distillation-based feature extraction algorithm to enhance the model's ability to discriminate source speakers. Our method employs a pre-training ASV model as a teacher network and the SSV model as a student network, using bona fide speech to guide the learning process. However, the improvements were marginal, particularly on the development set, indicating the complexity and resource demands of fine-tuning the distillation parameters. Our findings underscore the inherent difficulties in SSV and highlight the need for further research to develop more effective solutions. Besides, our submission won fourth place in the 2024 Source Speaker Tracking Challenge (SSTC).

Index Terms— Source speaker verification, SSTC, anti-spoofing

1. INTRODUCTION

Automatic speaker verification (ASV) is a crucial biometric technique designed to verify a speaker's identity through speech samples [1, 2]. With advancements in deep neural networks, ASV systems have achieved significant progress and found widespread applications [3]. However, these systems are vulnerable to spoofing attacks [4]. To address this challenge, various competitions and countermeasures have been developed [5, 6, 7]. These countermeasures are typically designed to distinguish between bona fide and spoofed speech samples, thereby safeguarding ASV systems [8, 9].

However, those countermeasures do not consider the task of verifying the origin (source) speaker's identity hidden in the spoofed speech samples [10]. With the development of voice conversion (VC), it's easy to convert the source speaker's identity to another target speaker's while keeping the linguistic content unchanged [11]. Although VC offers great convenience, it also poses serious threats to privacy and security. The source speaker verification (SSV) system aims to mitigate these threats by verifying the identity of the source speaker. As indicated in [10], VC cannot entirely eliminate linguistic factors associated with the source speaker, which are retained in the spoofed speech samples. Therefore, deep neural networks can be used to capture the redundant source speaker feature, thereby verifying the source speaker's identity. Moreover, the 'Source Speaker Tracing Challenge' (SSTC) [1] has been organized to encourage the development of SSV systems.

To some extent, there are similarities between the SSV and ASV, i.e., both of them typically use deep neural networks to extract embeddings related to the speakers' identity. However, the VC algorithm partially modifies the intrinsic features of the source speaker in the speech, making it challenging for the SSV model to capture sufficient information from spoofed speech samples to model the source speaker's identity. Therefore, we design a distillation-based feature extraction algorithm to encourage the SSV model to infer more discriminative source speaker embeddings. In this algorithm, the ASV model pre-training on Librispeech [12] is employed as the teacher network, while the SSV model pre-training on VoxCeleb2 [13] is utilized as the student network. During the training stage, for each spoofed speech passed to the student network, we sample a corresponding bona fide speech sample spoken by the same source speaker as the input of the teacher network. Besides, the teacher network parameters are fixed during the distillation process.

The results evaluated on Test-4 demonstrate that the proposed algorithm achieves a 0.38% improvement compared to the model without distillation. However, due to the deadline and limited resources, we only submitted the results of the student model to SSTC 2024, where it ranked fourth in the challenge. Additionally, the experimental results highlight the complexity of the SSV task, underscoring the need for further in-depth research by the community. The contributions of this paper are as follows:

(1) We built a source speaker verification system and won

† The corresponding author.

[1]https://sstc-challenge.github.io

fourth place in the SSTC 2024 competition.

(2) We propose a distillation-based feature extraction algorithm to encourage the model to extract more discriminative source speaker embeddings.

(3) We conducted systematic experiments and analyzed the experimental results, which proved that the SSV task is a challenging task that requires researchers to conduct more in-depth research.

The remainder of the paper is organized as follows: Section 2 describes our system submitted to the SSTC 2024 and presents the description of the proposed distillation-based feature extraction algorithm. The experimental setup and implementation are detailed in Section 3. Section 4 displays the experimental results. Finally, our conclusions and future work are provided in Section 5.

2. METHODS

In this section, we present the methodology employed for the SSV task. Our approach integrates the ECAPA-TDNN architecture [14], the Additive Angular Margin Softmax (AAM-Softmax) loss function [15], and a distillation-based feature extraction algorithm. This combination aims to enhance the discriminative power and generalization capability of the SSV model.

2.1. ECAPA-TDNN architecture

The ECAPA-TDNN is a state-of-the-art neural network architecture specifically designed for speaker verification tasks. It extends the traditional Time Delay Neural Network (TDNN) [16] by incorporating advanced techniques to improve performance.

As shown in Fig. 1, the core architecture of the ECAPA-TDNN consists of several convolutional layers with residual connections. These layers are designed to process the input speech features and extract high-level speaker-specific embeddings. The architecture typically includes:

- **TDNN Layers**: A series of TDNN layers with varying temporal contexts to capture different levels of temporal information.
- **SE-Res2Net Blocks**: These blocks incorporate Squeeze-and-Excitation (SE) and Res2Net modules to enhance feature representation.
- **Global Context Layer**: A layer to aggregate the global context from the feature maps.
- **Dense Layers**: Fully connected layers to refine the speaker embeddings.

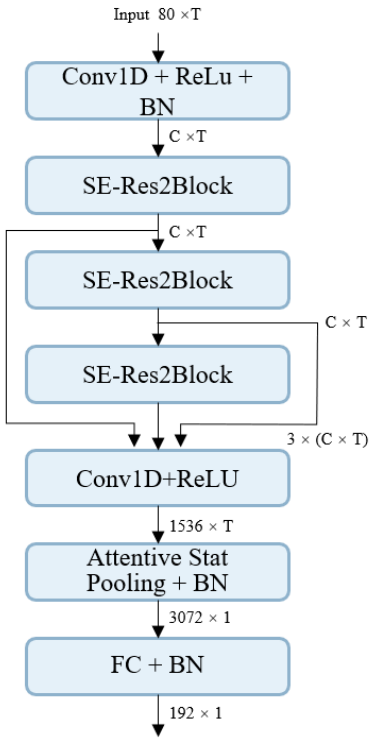

Fig. 1. Architecture of the ECAPA-TDNN. C and T denote the channel and temporal dimension of the intermediate feature map.

The global context layer in ECAPA-TDNN involves the use of intermediate feature maps at multiple stages of the network. This approach helps in capturing both local and global temporal patterns in the speech signal, which are crucial for robust speaker verification.

To train the ECAPA-TDNN network for the SSV task, we use the AAM-Softmax loss function. AAM-Softmax is an advanced loss function that enhances the discriminative power of speaker embeddings by introducing an angular margin between different classes (speakers). The loss function is formulated as follows:

$$\mathcal{L}_{AAM} = -\sum_{i=1}^{B} \log \frac{e^{s\cdot(\cos(\theta_{y_i}+m))}}{e^{s\cdot(\cos(\theta_{y_i}+m))} + \sum_{j\neq y_i}^{N} e^{s\cdot\cos(\theta_j)}}, \quad (1)$$

where B represents the size of the mini-batch and θ_{y_i} denotes the angle between the weight vector of the correct class y_i and the speaker embedding of the i-th training sample. The term θ_j refers to the angle between the weight vector of class j and the speaker embedding of the i-th training sample. The hyperparameter m is the angular margin, which increases the decision margin between classes to improve model generalization. The scaling factor s controls the distribution of logits, aiding in the stabilization of the training process. Finally, N is the total number of classes in the classification task. The loss function aims to maximize the difference between the correct

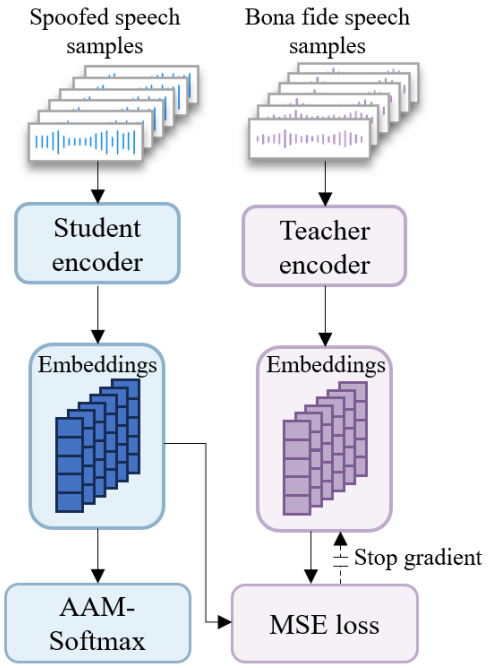

Fig. 2. Overview of the distillation-based feature extraction algorithm.

class and the incorrect classes by applying the margin m and scaling factor s to the cosine of the angles.

This formulation ensures that the network not only classifies the samples correctly but also maximizes the angular margin between different classes, leading to more discriminative speaker embeddings. The angular margin m ensures that embeddings for different speakers are well-separated, reducing intra-class variation and enhancing inter-class separation.

2.2. Distillation-based feature extraction

The distillation-based feature extraction algorithm is designed to leverage the knowledge from a pre-training ASV model to enhance the robustness and discriminative power of the SSV model. In this algorithm, the ASV model serves as the teacher network, while the SSV model acts as the student network. As shown in Fig. 2, the teacher network is trained on bona fide speech samples, which contain clean and authentic speaker characteristics, and the student network is trained on spoofed speech samples generated by VC algorithms. The objective is to align the embeddings produced by the student network with those produced by the teacher network, ensuring that the student network can infer robust speaker representations from the redundant source speaker feature in the spoofed speech.

During the training stage, spoofed speech samples are processed by the student network, while the corresponding bona fide speech samples from the same source speaker are passed to the teacher network. Additionally, the parameters of the teacher network remain fixed, providing a stable reference for the student network. The embeddings generated by the teacher network serve as targets of the student network for alignment purposes.

The alignment of the embeddings is achieved using the mean squared error (MSE) loss function [17]. The MSE loss measures the average squared difference between the embeddings from the teacher and student networks. It is defined as follows:

$$\mathcal{L}_{MSE} = \frac{1}{B}\sum_{i=1}^{B}\left\|\phi_T(x_i^{bona}) - \phi_S(x_i^{fake})\right\|^2, \quad (2)$$

where $\phi_T(x_i^{bona})$ denotes the embedding of the i-th bona fide speech sample x_i^{bona} produced by the teacher network, and $\phi_S(x_i^{fake})$ denotes the embedding of the corresponding spoofed speech sample x_i^{fake} produced by the student network. The MSE loss encourages the student network to produce embeddings that closely match those of the teacher network, thereby inferring more discriminative source speaker features.

The overall training objective of the student network is to minimize both the AAM-Softmax loss, which enhances the discriminative power of the embeddings, and the MSE loss, which ensures that the embeddings align with those produced by the teacher network. The total loss of the student network can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{AAM} + \lambda \cdot \mathcal{L}_{MSE}, \quad (3)$$

where λ denotes the distillation weight coefficient used to balance the contributions of the two loss functions. The combination of these loss functions enhances the SSV model's ability to identify the source speaker and improves its capability to capture the identity information of the source speaker within spoofed speech.

3. EXPERIMENTS

3.1. Database and evaluation metric

The SSV data set provided by the SSTC 2024 [18] is utilized for training our SSV models. Besides, we used the VoxCeleb2 development set [13] and the train-clean section of Librispeech [12] for training the pre-training ASV models. Following [18], the EER is adopted as our evaluation metric.

3.2. Training details

In this work, the ECAPA-TDNN with a 512-dimensional intermediate feature map is adopted as our backbone. The Adam optimizer is used to update the network parameters, with a cosine decay learning rate schedule. The initial learning rate is set to 1×10^{-3}, and the minimum learning rate is set to 1×10^{-5}. During the training stage, the speech samples are randomly cropped into four-second segments to form mini-batches and then used for extracting 80-dimensional

Table 1. Comparision of EER (%) between half small MFA-Conformer provided by SSTC 2024 and our ECAPA-TDNN under the development sets for SSV. The ECAPA-TDNN was trained on the data set mixed with all eight SSV training sets. 'AVE' donates the average EER over all development sets.

Trial lists	MFA-Conformer half small [18]	ECAPA-TDNN
Dev-1	9.40	11.79
Dev-2	8.62	10.21
Dev-3	7.67	8.66
Dev-4	7.59	9.77
Dev-5	7.51	8.69
Dev-6	12.89	13.52
Dev-7	32.48	27.47
Dev-8	28.80	26.05
Dev-9	34.05	33.24
Dev-10	45.77	40.38
Dev-11	17.21	16.43
Dev-12	20.81	20.51
AVE	19.40	18.89

Fbank features as the inputs for the models. Our models are trained for 25 epochs using an NVIDIA RTX 3090 GPU.

For the model trained on the data set mixed with all eight SSV training sets, the hyperparameters m and s in Eq. (1) are set to 0.15 and 32, respectively. Additionally, the network parameters are initialized with a model pre-training on the VoxCeleb2 development set to enhance performance.

For the model trained on the Train-4 data set, the hyperparameters m and s in Eq. (1) are set to 0.1 and 32, respectively. In the distillation procedure, the student network parameters are initialized with the model pre-training on the VoxCeleb2 development set, while the teacher network parameters are initialized with the model pre-training on the train-clean section of Librispeech.

4. RESULT

Table 1 and 2 compare the performance of our ECAPA-TDNN model with the half small MFA-Conformer system [18] provided by SSTC on the development and test sets, respectively. The results in Table 1 indicate that while the ECAPA-TDNN generally underperforms compared to the half small MFA-Conformer in seen VC scenarios, it excels in unseen VC scenarios. Specifically, on Dev-10, our model shows a performance improvement of over 11%, demonstrating superior generalization capabilities. Notably, the ECAPA-TDNN outperforms the half small MFA-Conformer in challenging seen VC scenarios such as Dev-7 and Dev-8.

Similarly, Table 2 presents results that reinforce this trend. Our ECAPA-TDNN model consistently outperforms the half small MFA-Conformer system in nearly all unseen VC scenarios (except Test-15). Moreover, it demonstrates significantly better performance in challenging seen VC scenarios (Dev-7, Dev-8, Test-7, and Test-8). Notably, under Test-7, our model achieves an improvement of more than 16%.

Table 2. Comparision of EER (%) between half small MFA-Conformer provided by SSTC 2024 and our ECAPA-TDNN under the test sets for SSV. 'AVE' donates the average EER over all testing sets.

Trial lists	MFA-Conformer half small [18]	ECAPA-TDNN
Test-1	9.79	12.71
Test-2	10.65	11.87
Test-3	7.00	8.62
Test-4	7.61	10.06
Test-5	6.73	8.59
Test-6	10.76	12.00
Test-7	32.90	27.63
Test-8	29.30	26.47
Test-9	34.59	33.95
Test-10	45.42	39.91
Test-11	18.71	17.54
Test-12	22.50	21.19
Test-13	36.66	35.57
Test-14	20.37	20.15
Test-15	9.53	10.24
Test-16	27.31	23.95
AVE	20.61	20.03

Furthermore, DET curves for both seen and unseen VC scenarios are illustrated in Fig. 3 and Fig. 4, respectively. These curves visually demonstrate that while our model outperforms the half small MFA-Conformer system in unseen VC scenarios, a substantial performance gap remains between seen and unseen VC results. This indicates that there is considerable potential for further improvement in the SSV task to enhance the model's generalization to unseen VC algorithms.

For the proposed distillation-based feature extraction algorithm, we compare its performance with the ECAPA-TDNN model across different values of the distillation weight parameter λ. The performances are evaluated on the Dev-4 and Test-4 data sets, with all models trained solely on the Train-4 training set. As shown in Table 3, the ECAPA-TDNN model without pre-training achieves an EER of 7.79% on Dev-4 and 7.42% on Test-4. the ECAPA-TDNN model pre-training on the VoxCeleb2 data set significantly enhances performance, reducing the EER to 7.10% on Dev-4 and 6.89% on Test-4. This demonstrates the efficacy of the pretraining model from large-scale datasets in improving SSV model performance.

When integrating the distillation-based feature extraction algorithm with the pre-training ECAPA-TDNN model, we observe varying impacts on the performance depending on the

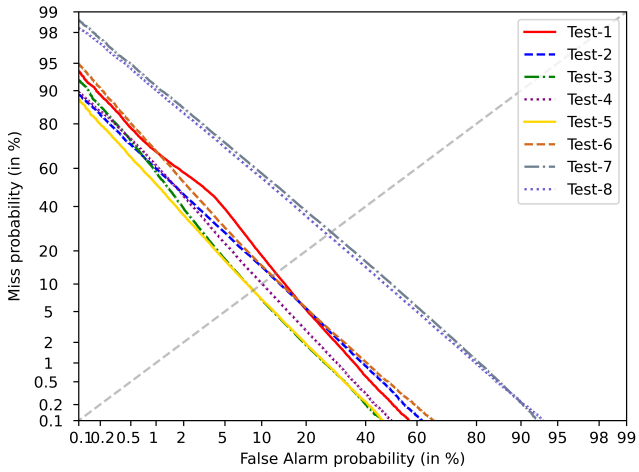

Fig. 3. DET curves of the ECAPA-TDNN for the seen VC scenarios, which shows the results from Test-1 to Test-8.

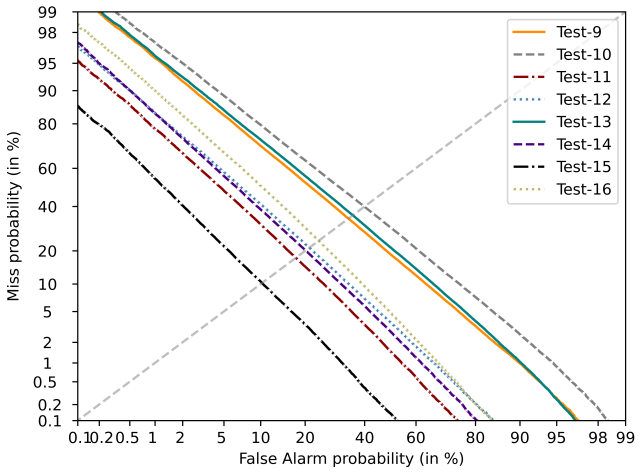

Fig. 4. DET curves of the ECAPA-TDNN for the unseen VC scenarios, which shows the results from Test-9 to Test-16.

Table 3. Comparison of performance between the ECAPA-TDNN and our proposed distillation-based feature extraction algorithm. All of them were trained only with the training set Train-4. The results in the table are the average of the EER (%) for the last three epochs.

Model	Parameters	Dev-4	Test-4
ECAPA-TDNN	6.65M	7.79	7.42
+ VoxCeleb2 pre-training		**7.10**	6.89
+ Distillation ($\lambda = 0.01$)	6.65M	7.20	6.83
Distillation ($\lambda = 0.1$)		7.14	6.66
Distillation ($\lambda = 0.3$)		7.24	6.64
Distillation ($\lambda = 0.5$)		**7.11**	**6.51**
Distillation ($\lambda = 0.7$)		7.34	6.65
Distillation ($\lambda = 1.0$)		7.22	6.64

value of the distillation weight parameter λ. All configurations lag behind the performance of the pre-training ECAPA-TDNN on the Dev-4 set, with results approaching the pre-training model's performance only at $\lambda = 0.5$. Nevertheless, the proposed algorithm outperforms the pre-training ECAPA-TDNN on the Test-4 set. The best performance is observed with $\lambda = 0.5$, where the EER decreases from 6.89% to 6.51%, indicating that a moderate amount of distillation enhances the model's discrimination ability.

Despite the improvement on Test-4, the performance is sensitive to the distillation weight coefficient, requiring careful adjustment, which is resource-intensive. Therefore, we only submitted the results of the student network to the SSTC2024 competition, achieving fourth place in this challenge.

Overall, these results highlight that SSV is a challenging task, and simple distillation cannot effectively align the embeddings of bona fide and spoofed speech. Therefore, more in-depth research is needed to uncover the relationship between the source speaker identity information in spoofed and bona fide speech, leveraging this information to enhance SSV model performance.

5. CONCLUSION AND FUTURE WORK

In this study, we developed an SSV system and proposed a novel distillation-based feature extraction algorithm that aims to enhance the discriminative power of the SSV model. Our experimental results indicate that the distillation-based feature extraction algorithm does not consistently improve performance. Although the best results were achieved with a moderate distillation weight parameter ($\lambda = 0.5$), the approach is sensitive to the distillation weight coefficient and resource-intensive. Consequently, we submitted only the student model in the proposed algorithm to the SSTC 2024 competition, where it achieved fourth place.

These findings highlight the inherent challenges in source speaker verification, particularly when dealing with spoofed speech generated by unseen voice conversion techniques. The complexity of capturing residual source speaker information from spoofed samples necessitates further research and development. In the future, we will focus on refining the distillation process, exploring alternative methods to enhance feature extraction, and conducting more comprehensive experiments to validate our methods to better understand the SSV task.

6. REFERENCES

[1] Xing Fan and John HL Hansen, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2010.

[2] Ruiteng Zhang, Jianguo Wei, Wenhuan Lu, Longbiao

Wang, Meng Liu, Lin Zhang, Jiayu Jin, and Junhai Xu, "ARET: Aggregated Residual Extended Time-Delay Neural Networks for Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 946–950.

[3] Ruiteng Zhang, Jianguo Wei, Wenhuan Lu, Lin Zhang, Yantao Ji, Junhai Xu, and Xugang Lu, "Cs-rep: Making speaker verification networks embracing re-parameterization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7082–7086.

[4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[5] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015, pp. 2037–2041.

[6] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[7] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[8] Anssi Kanervisto, Ville Hautamäki, Tomi Kinnunen, and Junichi Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 477–488, 2022.

[9] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023.

[10] Danwei Cai, Zexin Cai, and Ming Li, "Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] Bo Chen, Zhihang Xu, and Kai Yu, "Data augmentation based non-parallel voice conversion with frame-level speaker disentangler," *Speech Communication*, vol. 136, pp. 14–22, 2022.

[12] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[13] J Chung, A Nagrani, and A Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, 2018.

[14] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] Eric Bauer and Ron Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, vol. 36, pp. 105–139, 1999.

[18] Ze Li, Yuke Lin, Tian Yao, Hongbin Suo, and Ming Li, "The database and benchmark for source speaker verification against voice conversion," 2024.