Memory-Efficient Training for Text-Dependent SV with Independent Pre-trained Models

Seyed Ali Farokh
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
alifarokh@aut.ac.ir

Hossein Zeinali
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
hzeinali@aut.ac.ir

Abstract

This paper presents our submission to the Iranian division of the Text-Dependent Speaker Verification Challenge (TdSV) 2024. Conventional TdSV approaches typically jointly model speaker and linguistic features, requiring unsegmented inputs during training and incurring high computational costs. Additionally, these methods often fine-tune large-scale pre-trained speaker embedding models on the target domain dataset, which may compromise the pre-trained models' original ability to capture speaker-specific characteristics. To overcome these limitations, we employ a TdSV system that utilizes two pre-trained models independently and demonstrate that, by leveraging pre-trained models with targeted domain adaptation, competitive results can be achieved while avoiding the substantial computational costs associated with joint fine-tuning on unsegmented inputs in conventional approaches. Our best system reached a MinDCF of 0.0358 on the evaluation subset and secured first place in the challenge.

Keywords: Text-dependent Speaker Verification, Speaker Verification, Memory-efficient Training, Pre-trained Models, Transfer Learning

1 Introduction

Speaker verification (SV) is the task of confirming an individual's identity based on their voice. It involves comparing one or more enrollment utterances with a test utterance and can be performed in either a text-independent (TiSV) or text-dependent (TdSV) setting. In TiSV, the phonetic content of the utterances is unrestricted, and only the speaker's identity is verified, whereas in TdSV, the system verifies both the speaker's identity and the specific phrase spoken. With the development of various neural network architectures (Xie et al., 2019; Desplanques et al., 2020; Zeinali et al., 2019b; Snyder et al., 2018), loss functions (Xiang et al., 2019; Zhang and Koishida, 2017; Wang et al., 2018; Deng et al., 2019), and pooling methods (Snyder et al., 2018; India et al., 2019; Zhu et al., 2018), TiSV has seen considerable improvement in recent years, whereas TdSV has remained relatively underexplored. TdSV systems can be either phrase-dependent (i.e., shared passphrases), where a fixed set of phrases is predefined by the system, or phrase-independent (i.e., user-defined passphrases), allowing users to customize their phrases (Hossein et al., 2024). With the growing demand for voice-based authentication, TdSV has gained increasing attention, as the phonetic content can be used as passphrases (Tu et al., 2022), adding an extra layer of security to voice-based access control systems.

This paper presents our system submitted to Task 1 of the Text-dependent Speaker Verification Challenge 2024[1] (Zeinali et al., 2025), which aimed to encourage participants to explore novel approaches for TdSV. The challenge was organized into two divisions: an international one, which included two subtasks focusing on shared and user-defined passphrases, and an Iranian division, which mirrored Task 1 of the worldwide challenge but specifically emphasized developing methods with limited GPU resources. In this challenge, model enrollment is done using three enrollment utterances, and each trial consists of a test utterance and a model identifier. Speaker verification trials fall into one of the following categories:

- **Target Correct (TC)**: The speaker matches the claimed model and utters the correct phrase.
- **Target Wrong (TW)**: The speaker matches the claimed model but utters an incorrect phrase.

[1] Challenge website: https://tdsvc.github.io

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 95–102
November 20-22, 2025 ©2025 Association for Computational Linguistics

- **Impostor Correct (IC)**: The speaker does not match the claimed model but utters the correct phrase.

- **Impostor Wrong (IW)**: The speaker does not match the claimed model and utters an incorrect phrase. This category was excluded from the current year's challenge, as it does not pose sufficient difficulty for contemporary models.

In the context of TdSV, proposed systems are required to integrate both speaker and phrase verification scores and accept only TC trials[2]. Task 1 is phrase-dependent, employing a fixed set of ten phrases (five in Persian and five in English) for enrollment and testing. Additionally, to enhance the complexity of the challenge, some test utterances in TW trials were sourced from free-text recordings.

The primary evaluation metric adopted by TdSV 2024 is the normalized minimum Detection Cost Function (MinDCF), as defined in NIST SRE 2008 as a weighted sum of miss and false error probabilities, with $P_{target} = 0.01$, $C_{FalseAlarm} = 1$, and $C_{Miss} = 10$. The Equal Error Rate (EER) will also be reported as a secondary performance measure.

Previous successful approaches to TdSV typically jointly model speaker characteristics and the linguistic content of utterances. For instance, Liu et al. (2021) proposed a phoneme-aware attentive pooling method that incorporates frame-level phoneme posteriors into attentive pooling, improving the model's ability to utilize phonetic information effectively. Also, some studies have employed supervised multi-task learning to jointly learn speaker and linguistic features for further improvement (Yang et al., 2020; Han et al., 2021).

However, joint speaker and phrase modeling has some drawbacks compared to independent modeling. First, model development becomes more complex than developing the system based on independent phrase and speaker embedding models. Additionally, since phrase modeling requires attending to an entire utterance, inputs cannot be chunked during training, requiring variable-length inputs to be zero-padded. This issue substantially increases GPU memory requirements, particularly for recent transformer-based models, due to their quadratic time and memory complexity (Vaswani et al., 2017).

Furthermore, as demonstrated in this work, pre-trained speaker embedding models are highly effective at extracting speaker-related features while disregarding other information in input utterances. However, when subjected to multi-task fine-tuning, these models are prone to lose their initial ability to extract speaker-related features, allocating capacity to learning linguistic content instead. This shift reduces their effectiveness, especially when in-domain data for multi-task fine-tuning is limited.

Motivated by these challenges, we leverage the full capacity of pre-trained models and develop a TdSV system based on independent pre-trained models for phrase and speaker verification. For phrase verification, we fine-tune a pre-trained cross-lingual speech representation model for bilingual automatic speech recognition (ASR) in Persian and English, followed by a further fine-tuning stage for phrase classification. This classifier is used to reject incorrect phrases. Similarly, we develop several speaker embedding extractors based on pre-trained ResNets and Whisper (Radford et al., 2023) for our speaker verification system. After rejecting incorrect phrases using the phrase classifier, final verification scores are obtained by computing cosine similarity between test and enrollment embeddings.

Experimental results demonstrate that with well-designed fine-tuning stages, our TdSV system built on independently pre-trained models can achieve performance comparable to systems that jointly model speaker-related and linguistic information while using only a single Nvidia RTX 3090 GPU. This strategy substantially lowers GPU memory requirements and, consequently, reduces computational costs compared to the multi-GPU setups typically employed for training speaker recognition models (Zheng et al., 2023). Our best system secured first place in the Iranian division of the challenge and outperformed the third-place team in the international division (Zeinali et al., 2025).

The rest of the paper is organized as follows: Section 2 introduces the datasets used in this work. Sections 3 and 4 describe the architecture of our phrase and speaker verification systems, respectively. The experimental results and discussion are given in Section 5, and we conclude in Section 6.

2 Challenge Datasets

The DeepMine dataset (Zeinali et al., 2018, 2019a) is the primary source of the training and evalua-

[2] For Text-independent Speaker Verification (TiSV), the task definition differs: both TC and TW trials are accepted.

tion data for TdSV 2024. It was collected through crowd-sourcing, and while all participants were native Persian speakers, most contributed to the English portion of the dataset as well. The official TdSV 2024 data for Task 1 includes three subsets: training, development, and evaluation. The training subset consists of 183,431 utterances from 1,620 speakers. Among the utterances, 31,738 are free-text, while the rest were drawn from a fixed set of ten phrases comprising five Persian and five English phrases. The development and evaluation subsets are intended solely for system evaluation and contain 117,348 and 6,464,241 trials, respectively. During evaluation, model enrollment is conducted using three recordings of a specific phrase, and each trial includes a test utterance and a model identifier. The development set is provided to participants for evaluation and parameter tuning before submitting results to the official leaderboard. The evaluation subset is used for the official evaluation of the challenge. In addition to the DeepMine dataset, participants are also allowed to use the following datasets:

- **VoxCeleb 1&2** (Nagrani et al., 2017; Chung et al., 2018) are two large-scale datasets collected from YouTube videos, which contain over one million recordings from 7,205 celebrities. In this work, due to resource constraints, only VoxCeleb 1 was used, which includes over 100,000 utterances from 1,251 speakers.

- **LibriSpeech** (Panayotov et al., 2015) is a standard ASR corpus in US English that comprises approximately 1,000 hours of speech from 2,338 speakers. We only used the *train-clean-100* subset of this dataset to train our phrase verification system, which contains about 100 hours of speech.

- **Common Voice** (Ardila et al., 2020) is a multilingual speech dataset created from contributions of volunteers from worldwide. For this challenge, teams are restricted to using the Persian (Farsi) subset, which contains approximately 363 hours of validated speech from 4,148 speakers[3]. To prepare this subset for training our speaker verification systems, we excluded speakers with fewer than 30 recordings. From the remaining speakers with more than 650 recordings, we randomly selected 650 utterances per speaker, resulting in a final dataset with 125,017 utterances from 813 speakers.

The challenge rules prohibit the use of any other public or private data for training.

2.1 Data Augmentation

We did not use any augmentation methods in our phrase verification system. However, following the previous successful studies on speaker verification (Chen et al., 2022; Zheng et al., 2023), we adopted SoX-based speed perturbation by factors of 0.9 and 1.1 to triple the number of speakers during training, followed by an on-the-fly implementation of the following augmentations, each applied with a probability of 0.6: noise addition using the MUSAN dataset (Snyder et al., 2015), reverberation using RIRs dataset (Ko et al., 2017), and gain augmentation.

3 Phrase Verification System

Our proposed system for TdSV 2024 consists of two independent subsystems for phrase and speaker verification. The phrase verification system is a classifier that rejects TW trials, while the speaker verification system is responsible for producing similarity scores. Although this system design does not benefit from joint modeling of speaker and text, it greatly simplifies the system development process and allows for the use of various pre-trained models for each subsystem with minimal modifications.

The phrase classifier is an 11-class model trained with standard softmax. The first ten classes correspond to the set of phrases in the challenge, and the final class represents free text (or "none of the above"). This classifier is built on XLSR[4] (Conneau et al., 2021), a pre-trained cross-lingual speech representation model trained by solving a self-supervised contrastive task, proven to be effective in low-resource languages compared to traditional feature extraction methods. This model takes a raw waveform as input and produces a sequence of features.

Moreover, to improve the model's ability to extract linguistic features from Persian and English inputs, we initially fine-tuned the XLSR for bilingual speech recognition in Persian and English.

[3] Common Voice 18.0, released on 6/19/2024

[4] Facebook/wav2vec2-xls-r-300m

System	Full Training			Domain Adaptation		
	Epoch	BS	LR	Epoch	BS	LR
S2	-	-	-	15	32	3e-4
S3	100	64	1e-3	15	32	3e-4
S4	15	64	1e-3	7	28	5e-5
S5	15	64	1e-3	7	28	5e-5

Table 1: Hyper-parameters used in different submitted systems S2–S5 (BS = batch size, LR = learning rate).

During this phase, 30% of the training subset of Common Voice Farsi and LibriSpeech (*train-clean-100*) were used, and the model was trained using CTC loss (Graves et al., 2006) for 40 epochs, with an initial learning rate of 0.001 and an effective batch size of 32. In our experiments, this phase contributes to improving the performance of the phrase verification system.

Finally, to train the classifier, an attention-based pooling layer was added to the fine-tuned XLSR to compute fixed-dimensional utterance-level feature vectors from frame-level representations h_t ($t = 1, ..., T$):

$$e_t = W_1 h_t + b_1, \tag{1}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau}^{T} \exp(e_\tau)}, \tag{2}$$

$$\tilde{h} = \sum_{t}^{T} \alpha_t (W_2 h_t + b_2), \tag{3}$$

where, e_t and α_t are the attention score and weight, respectively. $\tilde{h}$ refers to the utterance-level feature vector, which is finally fed to a fully connected layer with ReLU activation, followed by a linear classifier. The network was trained using the Cross-Entropy loss function for one epoch on the entire training samples of the challenge dataset, with a learning rate of 0.0005 and an effective batch size of 64.

4 Speaker Verification System

To leverage the full power of pre-trained SV models and mitigate the computational cost of training randomly initialized models, we explored two directions for developing our SV system. In the first approach, we fine-tuned several pre-trained ResNet-based models, widely used as a standard architecture in speaker verification. In the second approach, we studied the performance of pre-trained ASR models adapted for SV, which have shown promising results in previous studies (Zhang et al., 2022; Cai et al., 2023; Liao et al., 2023). More specifically, we employed the Whisper-PMFA (Zhao et al., 2024) method, which involves fine-tuning a pre-trained Whisper model for speaker recognition.

4.1 Training Protocol

We trained our models in two stages:

- **Full training** (T_1): In this stage, models were trained on a combination of out-of-domain data (Common Voice Farsi and VoxCeleb 1) and in-domain (DeepMine) data, totaling 3,684 speakers, to learn robust and generalizable speaker embeddings across different domains. Pre-trained ResNets did not undergo this stage, as they are already capable of extracting rich speaker-specific features. During this phase, 300 consecutive frames of each input utterance were randomly selected in each epoch to prevent overfitting, reduce GPU memory usage, and accelerate training. Moreover, all augmentation methods explained in Section 2.1 were applied. We employed the widely used AAM-Softmax (Deng et al., 2019) loss with the subcenter method and the Inter-TopK penalty (Zhao et al., 2021) to train our models, with a constant margin and scale of 0.2 and 32, respectively.

- **Domain adaptation** (T_2): We fine-tuned our models using in-domain data after full training to bridge the domain gap and improve performance. During this stage, augmentation methods and the Inter-TopK penalty were removed to prevent domain mismatch. Additionally, the number of randomly selected frames was increased from 300 to 600 to enhance the models' generalization capability (Garcia-Romero et al., 2019, 2020). Fine-tuning was performed with smaller learning rates to preserve the models' generalization abilities.

All models were optimized using SGD with a momentum of 0.9 and a weight decay of 1e-4. We also utilized an exponential decay scheduler with a minimum learning rate of 5e-5 for T_1 and 1e-6 for T_2. Other training hyper-parameters are listed in Table 1. Note that gradient accumulation was used to achieve the target effective batch size when GPU memory was limited. The dimensionality of speaker embeddings was set to 256 across all models. All experiments were conducted on a sin-

System	Architecture	Training Stages	Development		Evaluation	
			MinDCF$_{0.01}$	EER(%)	MinDCF$_{0.01}$	EER(%)
S1	ResNet34		0.0614	1.3938	0.0784	1.7390
S2	ResNet293	T_2	0.0225	0.8733	**0.0376**	**1.1080**
S3	ResNet152	$T_1 + T_2$	0.0191	0.6757	0.0764	2.3444
S4	Whisper-PMFA	$T_1 + T_2$	0.0163	**0.6121**	0.0584	2.0410
S5	Whisper-PMFA	$T_1 + T_2$	**0.0161**	0.6126	0.0583	2.0445
Fusion (S1∼S5)			**0.0119**	**0.5605**	**0.0358**	1.2457

Table 2: Results of different submissions on the development and evaluation sets.

Subset	MinDCF$_{0.01}$	EER(%)
Development	0.0000	0.00
Evaluation	0.0003	0.01

Table 3: Phrase verification performance on TC-vs-TW trials.

gle Nvidia RTX 3090 GPU using the WeSpeaker toolkit (Wang et al., 2024).

4.2 ResNet

ResNet (Xie et al., 2019) is a widely used architecture for speaker recognition that has performed excellently in previous speaker verification challenges (Zheng et al., 2023). Consequently, many open-source implementations and pre-trained models have been publicly released based on this architecture. Trained on large-scale datasets like VoxCeleb 1&2, these pre-trained models can provide a robust starting point for training speaker recognition models on other datasets by improving their generalization and speeding up the convergence.

During the challenge period, we submitted three systems based on a bottleneck-block ResNet, all adopting temporal statistics pooling (Snyder et al., 2018) for aggregating variable-length sequence features into utterance-level embeddings. The first system (S1) was a pre-trained ResNet34 without domain adaptation, while the second one (S2) was a pre-trained ResNet293 that underwent domain adaptation. Finally, we applied both training stages to a randomly initialized ResNet152 to obtain our last ResNet-based system (S3).

4.3 Whisper-PMFA

Building on the successful use of pre-trained ASR models in speaker verification (Zhang et al., 2022; Cai et al., 2023), Zhao et al. (2024) recently proposed Whisper-PMFA (Partial Multi-Scale Feature Aggregation using Whisper) to leverage the capabilities of Whisper, a large-scale multilingual ASR model based on transformer architecture. Whisper-PMFA adapts Whisper for speaker verification by selectively concatenating frame-level outputs from specific transformer layers rather than aggregating features from all layers. This approach not only reduces computational overhead but also enhances performance by minimizing the integration of irrelevant information from lower-impact layers.

Inspired by this, we studied the performance of Whisper-PMFA in this challenge. Since Whisper was not trained for the speaker recognition task, we applied both training stages to Whisper-PMFA. Additionally, before the full training stage, we froze the Whisper parameters and fine-tuned the model for five epochs to prevent updating the pre-trained model in the wrong direction due to the random initialization of newly added components. We submitted two Whisper-PMFA-based systems (S4 and S5) to this challenge, differing only in the AAM-Softmax margin used during the domain adaptation phase: 0.35 for S4 and 0.2 for S5.

4.4 Feature Extraction

80-dimensional log Mel filter bank energies with a 25ms window and 10ms frame-shift were extracted for our ResNet-based models. Voice activity detection (VAD) was not applied, and all features were mean-normalized. Likewise, 80-dimensional log magnitude Mel spectrograms consistent with the pre-trained Whisper were utilized for training Whisper-PMFA.

4.5 Backend

Speaker embeddings were extracted from the final fully connected layer of the models, and cosine similarity was used to compute scores. Since model

Methods	Development	
	MinDCF$_{0.01}$	EER(%)
Whisper-PMFA (T$_1$)	0.0234	0.9253
+ Domain adaptation (T$_2$)	0.0177	0.6273
++ AS-Norm	0.0161	0.6126

Table 4: Ablation study on Whisper-PMFA.

enrollment is done using three utterances in this challenge, we used the average of embedding vectors of each model during scoring.

Afterward, AS-Norm (Wang et al., 2020) was used for score normalization, using 1,620 cohorts obtained from speaker-wise averaging of all embeddings in the training subset of the challenge dataset. The top 300 most similar scores were selected to compute the mean and standard deviation for normalization.

Finally, we adopted score fusion by averaging single-system scores to further improve performance.

5 Results

Table 2 shows the evaluation results of our single and fusion systems on the development and evaluation subsets of the challenge after applying AS-Norm and rejecting TW trials. The results indicate that the Whisper-PMFA method outperforms the widely used ResNet architecture with random initialization, conforming to the findings of previous studies on the effectiveness of adapting pre-trained ASR models for speaker verification. However, it can be observed from the results that the ResNets pre-trained on approximately twice the data (VoxCeleb 1&2) can considerably surpass Whisper-PMFA after a well-designed domain adaptation stage, which highlights the importance of large-scale pre-training in improving the generalization ability of speaker verification models.

In addition, Figure 1 presents the Detection Error Tradeoff (DET) curves of the best-performing system for different categories of evaluation data. The results indicate that the model generally performs better on Persian phrases, which is expected given that the DeepMine dataset was collected from native Persian speakers, many of whom are likely less fluent in English. Furthermore, the results show noticeably higher performance for male speakers compared to female speakers. This disparity is not solely due to the inherent challenges of verifying female voices, but is also influenced by the specific characteristics of the DeepMine dataset, as discussed in its original description (Zeinali et al., 2018, 2019a) and in the official challenge results paper (Zeinali et al., 2025).

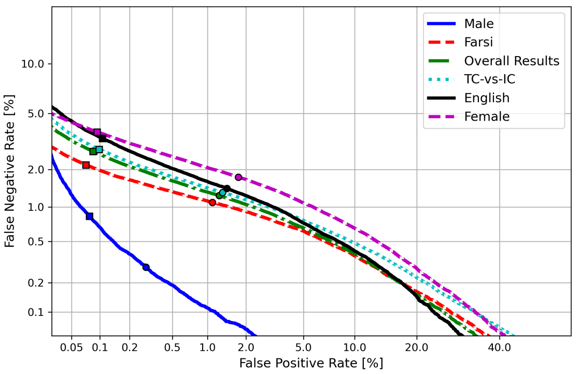

Figure 1: DET curves of our best-performing system.

We also report the MinDCF and EER of the proposed phrase verification system on TC-vs-TW trials of the development and evaluation subsets (Table 3). According to the results, our phrase verification system demonstrates a near-optimal performance on this task.

5.1 Ablation Study

We conducted an ablation study on our Whisper-PMFA system (S5). The development set of the challenge dataset was used as our evaluation benchmark. We can observe from the results (Table 4) that the domain adaptation phase improved the MinDCF from 0.0234 to 0.0177. Also, a further improvement of MinDCF to 0.0161 was achieved after applying AS-Norm.

5.2 Comparison with Other Teams

To contextualize our performance, we report in Table 5 the evaluation results of our best system alongside the top-performing submissions in Task 1 of the international division of the TdSV Challenge. Team names and scores are taken directly from the official challenge results paper (Zeinali et al., 2025), which also provides brief descriptions and comparisons of the proposed architectures. As shown, our system achieves a lower MinDCF than the team ranked third in the international division.

6 Conclusion

In this paper, we present our system for Task 1 of the Iranian division of the Text-dependent Speaker Verification (TdSV) Challenge 2024, fo-

Team	MinDCF$_{0.01}$	EER(%)
Team 04 (Sreekanth, 2024)	**0.0297**	1.132
Team 08	0.0326	**1.013**
Our System	0.0358	1.246
Team 02	0.0379	1.164
Team 01	0.0504	2.245

Table 5: Evaluation results for our best system and the top-ranked teams in Task 1 of the international division of TdSV.

cusing on resource-constrained training for TdSV systems. Unlike previous methods that jointly model speaker-related and linguistic features, our approach leverages two independent pre-trained models for phrase and speaker verification. This design reduces the computational costs associated with joint modeling during training while fully utilizing the capabilities of pre-trained models to achieve competitive performance. Our best system achieved a MinDCF of 0.0358 on the evaluation subset, securing first place in the challenge.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Danwei Cai, Weiqing Wang, Ming Li, Rui Xia, and Chuanzeng Huang. 2023. Pretraining Conformer with ASR for speaker verification. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, and Yanmin Qian. 2022. SJTU-AISpeech system for VoxCeleb speaker recognition challenge 2022. *arXiv preprint arXiv:2209.09076*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. In *Interspeech 2018*. ISCA.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Daniel Garcia-Romero, Greg Sell, and Alan Mccree. 2020. MagNetO: X-vector magnitude estimation network plus offset for improved speaker recognition. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 1–8.

Daniel Garcia-Romero, David Snyder, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. 2019. X-Vector DNN refinement with full-length recordings for speaker recognition. In *Interspeech 2019*, pages 1493–1496.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Bing Han, Zhengyang Chen, Zhikai Zhou, and Yanmin Qian. 2021. The SJTU system for short-duration speaker verification challenge 2021. In *Interspeech 2021*, pages 2332–2336.

Zeinali Hossein, Lee Kong Aik, Alam Jahangir, and Burget Lukas. 2024. Text-dependent speaker verification (TdSV) challenge 2024: Challenge evaluation plan. *arXiv preprint arXiv:2404.13428*.

Miquel India, Pooyan Safari, and Javier Hernando. 2019. Self multi-head attention for speaker recognition. In *Interspeech 2019*, pages 4305–4309.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5220–5224. IEEE.

Dexin Liao, Tao Jiang, Feng Wang, Lin Li, and Qingyang Hong. 2023. Towards a unified Conformer structure: from ASR to ASV task. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yan Liu, Zheng Li, Lin Li, and Qingyang Hong. 2021. Phoneme-aware and channel-wise attentive learning for text dependent speaker verification. In *Interspeech 2021*, pages 101–105.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Sankala Sreekanth. 2024. Exploring self-supervised representations for text-dependent speaker verification. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1232–1239.

Youzhi Tu, Weiwei Lin, and Man-Wai Mak. 2022. A survey on text-dependent and text-independent speaker verification. *IEEE Access*, 10:99038–99049.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Shuai Wang, Zhengyang Chen, Bing Han, Hongji Wang, Chengdong Liang, Binbin Zhang, Xu Xiang, Wen Ding, Johan Rohdin, Anna Silnova, et al. 2024. Advancing speaker embedding learning: WeSpeaker toolkit for research and production. *Speech Communication*, 162:103104.

Weiqing Wang, Danwei Cai, Xiaoyi Qin, and Ming Li. 2020. The DKU-DukeECE systems for voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2010.12731*.

Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu. 2019. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1652–1656.

Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2019. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5791–5795. IEEE.

Yexin Yang, Shuai Wang, Xun Gong, Yanmin Qian, and Kai Yu. 2020. Text adaptation for speaker verification with speaker-text factorized embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6454–6458.

Hossein Zeinali, Lukáš Burget, and Jan Honza Černockỳ. 2019a. A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 397–402. IEEE.

Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukáš Burget. 2025. Text-dependent speaker verification challenge 2024: Exploring shared and user-defined passphrases. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hossein Zeinali, Hossein Sameti, and Themos Stafylakis. 2018. DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English. In *Odyssey*, pages 386–392.

Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. 2019b. BUT system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*.

Chunlei Zhang and Kazuhito Koishida. 2017. End-to-end text-independent speaker verification with triplet loss on short utterances. In *Interspeech 2017*, pages 1487–1491.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng. 2022. MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification. In *Interspeech 2022*, pages 306–310.

Miao Zhao, Yufeng Ma, Min Liu, and Minqiang Xu. 2021. The SpeakIn system for voxceleb speaker recognition challange 2021. *arXiv preprint arXiv:2109.01989*.

Yiyang Zhao, Shuai Wang, Guangzhi Sun, Zehua Chen, Chao Zhang, Mingxing Xu, and Thomas Fang Zheng. 2024. Whisper-PMFA: Partial multi-scale feature aggregation for speaker verification using Whisper models. In *Interspeech 2024*, pages 2680–2684.

Yu Zheng, Yajun Zhang, Chuanying Niu, Yibin Zhan, Yanhua Long, and Dongxing Xu. 2023. Unisound system for voxceleb speaker recognition challenge 2023. *arXiv preprint arXiv:2308.12526*.

Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. 2018. Self-attentive speaker embeddings for text-independent speaker verification. In *Interspeech 2018*, pages 3573–3577.