

# Text-dependent Speaker Verification Challenge 2024: Exploring Shared and User-defined Passphrases

Hossein Zeinali

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

Kong Aik Lee

Department of Electrical and Electronic Engineering  
The Hong Kong Polytechnic University  
Kowloon, Hong Kong  
kong-aik.lee@polyu.edu.hk

Jahangir Alam

Computer Research Institute of Montreal  
Montreal, Canada  
jahangir.alam@crim.ca

Lukáš Burget

Department of Computer Graphics and Multimedia  
Brno University of Technology  
Brno, Czech Republic  
burget@fit.vutbr.cz

**Abstract**—In contrast to text-independent speaker verification, which has received significant attention from researchers and has many competitions dedicated to it, text-dependent speaker verification (TdSV) has been less explored recently. The TdSV Challenge 2024 was organized to analyze and explore novel methods for this type of speaker verification and aims to motivate participants to develop new approaches to TdSV, conduct comprehensive analyses, and investigate advanced techniques such as self-supervised learning. This challenge builds on the achievements of the short-duration speaker verification (SdSV) Challenges held in 2020 and 2021 and focuses specifically on TdSV in two distinct scenarios. The first scenario involves conventional TdSV, while the second focuses on speaker enrollment using user-defined passphrases. This paper provides a detailed description of both tasks, introduces the evaluation rules, and presents a comprehensive analysis of the results obtained from this challenge.

**Index Terms**—Text-dependent Speaker Verification, TdSV Challenge, DeepMine dataset, User-defined passphrase

## I. INTRODUCTION

The rapidly evolving field of speaker verification has witnessed significant advancements, particularly in the context of text-independent speaker verification, due to availability of various large datasets [1], [2] and many valuable competitions [3]–[6]. The text-dependent speaker verification (TdSV) challenge 2024 aims to transfer these advancements to this area by facilitating the exploration of innovative approaches and methodologies to enhance the performance and robustness of TdSV systems. This challenge, through its structured and competitive format, aims to motivate participants to push the boundaries of current technologies and methodologies, particularly focusing on multi-task learning, self-supervised learning, and few-shot learning. These advanced techniques have shown promising results in various machine learning domains and are now being applied to tackle the unique challenges posed by text-dependent speaker verification.

The TdSV Challenge 2024 continues the successful Short-duration Speaker Verification (SdSV) Challenges held in 2020 and 2021 [7]. While the previous challenges encompassed a broader scope of speaker verification tasks, the 2024 challenge focuses on text-dependent scenarios. This specificity is intended to drive more concentrated efforts and innovations in this niche yet critical area.

The challenge committee thanks the Iran Vice-Presidency for Science, Technology, and Knowledge-Based Economy for sponsoring the challenge. The committee also acknowledges Sharif DeepMine for providing the challenge dataset.

Participants are requested to address two tasks within this challenge, each posing unique characteristics and requiring different approaches and strategies. In the first task, fixed passphrases are used, while the second task utilizes user-defined passphrases. The second version of the DeepMine dataset is used for both, which provides a rich and diverse set of data necessary for developing and testing robust speaker verification systems.

The comprehensive nature of the TdSV Challenge 2024, with its two distinct tasks, provides a ground for testing and refining various machine learning approaches in the context of TdSV. By leveraging the DeepMine dataset and focusing on both fixed and user-defined passphrases, this challenge offers a robust platform for participants to demonstrate the efficacy of their approaches. The outcomes of this challenge are expected to contribute significantly to the field, paving the way for more secure, efficient, and accurate speaker verification systems in the future.

## II. CHALLENGE TASKS AND DATASET

This section provides a brief description of the two tasks of the TdSV challenge. Additionally, the allowed datasets for this challenge are described. Further details about the challenge are available in the challenge description [8].

### A. Tasks Description

1) *Task 1: Conventional TdSV*: Task 1 of the TdSV Challenge entails performing speaker verification in the standard text-dependent mode, where the user's passphrase is selected from a pre-defined closed set of phrases. This task requires a dual verification process in which the speaker's identity and the target passphrase must be confirmed. Given a test speech segment and the enrollment data of the target speaker, the objective is to ascertain whether the target speaker uttered the test segment and the specific phrase. Each trial comprises a test segment and a model identifier that indicates a phrase ID and three enrollment utterances.

The enrollment and test phrases (except for target-wrong trials where some tests were selected from free-text utterances) are selected from a predetermined set of 10 sentences, which includes five Persian phrases and five English phrases (see Table I). Table II displays the number of trials categorized by language, gender, and trial type. In TdSV, there are four trial types: *target-speaker/correct-phrase*

TABLE I

PHRASES USED IN TASK 1 OF THE CHALLENGE. AMONG THEM, PHRASES 01, 04, 06, AND 08 ARE USED AS USER-DEFINED PASSPHRASES FOR TASK 2. THE TRANSLITERATIONS ARE PROVIDED FOR THE PERSIAN PHRASES.

Id	Phrase
01	sedAye man neshAndahandeye hoviyyate man ast.
02	sedAye har kas monhaser be fard ast.
03	hoviyyate man rA bA sedAye man tayid kon.
04	sedAye man ramze obure man ast.
05	baniAdam azAye yekdigarand.
06	My voice is my password.
07	OK Google.
08	Artificial intelligence is for real.
09	Actions speak louder than words.
10	There is no such thing as a free lunch.

TABLE II

NUMBER OF TRIALS IN EACH PARTITION FOR TASK 1.

Language	Gender	TC	TW	IC
Farsi	Male	89,739	278,958	1,250,792
Farsi	Female	161,211	480,052	1,482,490
English	Male	74,947	217,124	836,901
English	Female	136,905	393,263	1,061,859
Total		462,802	1,369,397	4,632,042

(TC), *target-speaker/wrong-phrase* (TW), *impostor-speaker/correct-phrase* (IC), and *impostor-speaker/wrong-phrase* (IW) [9]. However, IW trials are excluded from this challenge due to their ease of rejection. The systems should only accept TC trials and reject all other types (unlike text-independent speaker verification, TW trials must be rejected in this context).

2) *Task2 - TdSV Using User-defined Passphrase*: In Task 2, we explore a more complex scenario where speaker verification is based on user-defined passphrases. Unlike Task 1, there are no predefined phrases for training; users choose their own passphrases. This setup creates a realistic and challenging environment, requiring the system to adapt to previously unseen phrases during both enrollment and verification.

Each trial includes a test speech segment and a model identifier featuring three repetitions of the user-defined pass-phrase and several additional free-text (FT) enrollment utterances. Participants can use these FT utterances to improve the speaker model, allowing them to assess how extra data impacts system performance. The trial types are the same as in Task 1, where TC trials should be accepted, and TW and IC trials should be rejected. The number of trials for Task 2 is detailed in Table III.

To simulate user-defined passphrases, four out of the ten available text-dependent phrases in the dataset are used exclusively in the test set. At the same time, the remaining six are included in the training data. Consequently, the training data for this task consists of text-independent Persian utterances and utterances from these six phrases.

### B. Training and Evaluation Data

The evaluation and in-domain training data for the challenge are selected from a new version of the DeepMine dataset [10], [11]. For both tasks, a fixed training condition is used, where systems must be trained only with a designated set composed of VoxCeleb 1 and 2 [1], [2], LibriSpeech [12], Mozilla Common Voice Farsi [13], and task-specific in-domain training data, which includes utterances from **1620 speakers**, some with only Persian phrases. Model enrollment is conducted in a language-dependent manner, and there are no cross-language trials in the challenge. For each task, using the other task's

TABLE III

NUMBER OF TRIALS IN EACH PARTITION FOR TASK 2.

Language	Gender	TC	TW	IC
Farsi	Male	35,955	114,004	497,432
Farsi	Female	65,340	195,543	598,391
English	Male	30,250	63,875	331,392
English	Female	55,458	115,718	430,415
Total		187,003	489,140	1,857,630

in-domain training data is prohibited. This rule is crucial for Task 2 because the user-defined passphrase scenario was created using the same 10 text-dependent phrases in the DeepMine dataset.

This year, a *separate development set* was provided for the challenge. Teams could use this set for parameter tuning and system evaluation before submitting them to the leaderboard to minimize the number of submissions. Any other training, such as fusion training on the development set, was not permitted.

### C. Evaluation Metrics

The main performance metric adopted for the challenge is the normalized minimum *Detection Cost Function* (DCF) defined as a weighted sum of the miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}),$$

where,  $C_{Miss} = 10$ ,  $C_{FalseAlarm} = 1$  and  $P_{Target} = 0.01$ . Based on the parameters, the normalized DCF ( $DCF_{norm}$ ) is DCF divided by 0.1 as the best cost that could be obtained by constant decision (i.e. rejecting all trials). In addition to  $MinDCF_{norm}^{0.01}$ , the Equal Error Rate (EER) is reported as the common metric in speaker verification.

## III. CHALLENGE RESULTS

In the TdSV Challenge 2024, we received registrations from 20 teams across various countries. Among them, eight teams submitted at least one result for Task 1 and six teams for Task 2. This section first explains the contributions of the best-performing teams for both tasks, then presents the overall challenge results for Task 1 and 2. Subsequently, analyses of the selected teams' results will be provided.

### A. Best Performing Systems

1) *Team 02*: This team employed three large ResNet [14] encoders—ResNet221, ResNet293, and ResNet512—in Task 1 to extract speaker embeddings. The networks were initially trained with softmax loss and then fine-tuned with large-margin optimization [15]. Scoring was based on cosine similarity, and score normalization was done using Consistency Measure Factor (CMF) [16] and Adaptive Symmetric Normalization (AS-norm). For Task 2, a Wav2Vec2 [17] encoder was added to the pipeline to reject Target-Wrong trials.

2) *Team 04*: This team used a pre-trained WavLM model [18] to extract features from utterances. A Multi-Head Factorized Attentive Pooling (MHFA) layer [18] was applied to generate a fixed-size vector for each utterance. A segment-level encoder with dual softmax heads was used for speaker and speaker-phrase classification. A similar network was trained for phrase classification to reject incorrect trials. Scores were computed using cosine similarity.

TABLE IV

RESULTS OF CONVENTIONAL TdSV IN TASK 1. RESULTS ARE SORTED BASED ON THE MINDCF. # SUB. MEANS NUMBER OF SUBMISSIONS.

Team IDs	# Sub.	EER [%]	MinDCF <sup>0.01</sup> <sub>norm</sub>
Team 04	9	1.132	<b>0.0297</b>
Team 08	13	<b>1.013</b>	0.0326
Team 02	8	1.164	0.0379
Team 01	12	2.245	0.0504
Team 05	14	1.964	0.0633
Team 07	5	17.762	0.2279
Team 06	1	32.250	0.5337
Team 03	4	17.166	0.9420

TABLE V

DETAILED RESULTS FOR THE BEST PERFORMING TEAM (I.E. TEAM 04) IN TASK 1.

Condition	EER	MinDCF
Overall Results	1.13	0.0297
Male	0.32	0.0145
Female	1.58	0.0381
Farsi	1.00	0.0255
English	1.27	0.0341
TC-vs-IC	1.19	0.0316
TC-vs-TW	0.23	0.0053

3) *Team 05*: This team proposed using speaker embedding extractors like ResNet34 and ECAPA-TDNN [19], pre-trained on out-of-domain challenge datasets. In the second stage, large-margin fine-tuning was applied using the in-domain dataset. A speech recognition system was trained to filter incorrect trials, with scores in the backend computed using cosine similarity.

4) *Team 08*: This team similarly used a pre-trained large WavLM model to extract features from input speech. Then, three ECAPA-TDNN classifiers were trained on these extracted features. The first classifier identifies speakers, the second classifies phrases to reject Wrong Trials, and the output of the third classifier was the combined speaker-phrases labels. The embedding vectors extracted from these classifiers are used for scoring and score fusion. Results show that the best performance was achieved when all three classifiers were used together.

### B. Task 1 Overall Results

As mentioned earlier, we received submissions from eight teams. Table IV compares the results for these teams. The first team performs best based on DCF, while the second team has slightly better results based on EER. The top three teams show similar performance in terms of EER, but the differences based on DCF are more significant. The fourth and fifth teams also have comparable results. The remaining three teams have considerably worse results, indicating they struggle with different trial types.

1) *Detailed Results of Best Performing Team in Task 1*: To better analyze the results obtained from the challenge, detailed results of the first team for Task 1 (i.e., Team 04) are presented in this section. Table V shows the detailed results, and Figure 1 illustrates the DET curve for different conditions in this task.

As the results indicate, performance for the male gender is significantly better than for the female. This difference is not solely because speaker verification for women’s voices is more challenging but also depends on the DeepMine dataset. In this dataset, female speakers participated more, and several of them recorded using different mobile devices, making verification harder for females.

Comparing results for Persian and English languages shows that language difference does not significantly impact performance, al-

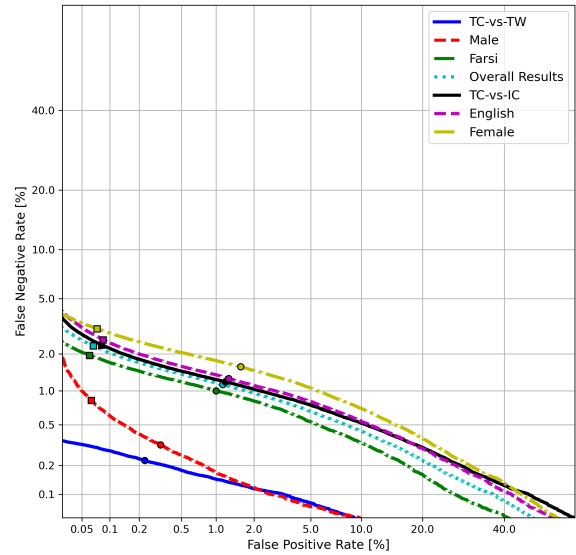


Fig. 1. DET curves of the best performing system for Task 1

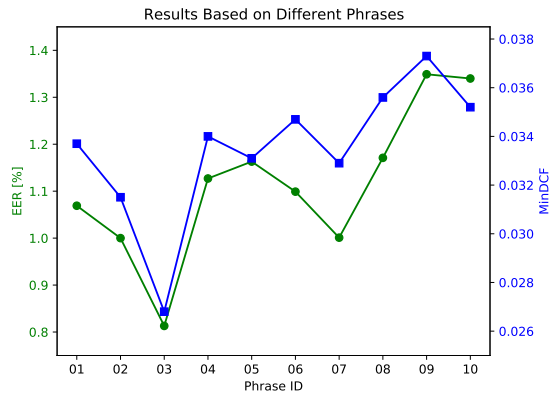


Fig. 2. The average results of different phrases for the top three teams in Task 1

though all DeepMine project participants were native Persian speakers with different English proficiency.

Finally, the result of TC-vs-TW tests shows that this team managed to reject TW trials quite well, but errors are still notable. For this type, Team 08 was able to detect and reject these tests with almost zero error. Therefore, improving this section could further enhance Team 04’s results.

It is worth noting that increasing the in-domain dataset for text-dependent speaker verification in the future is expected to lead to further improvements in results.

2) *Analysis of Results for Different Phrases in Task 1*: To further analyze the results, this section breaks down the performance metrics for each phrase in the dataset. Figure 2 illustrates these metrics for each phrase based on the average results of 3 top best-performing teams. Notably, a consistent trend is observed for both criteria.

Among the English phrases, “OK Google” demonstrates the best performance across both criteria despite being the shortest phrase in the dataset. This could be attributed to the native language of the participants. As a familiar and easy phrase, participants are likely to pronounce it without difficulty. Conversely, the other phrases pose more challenges for speakers with lower English proficiency.

Overall, no significant conclusions can be drawn from the results, except that there is no substantial performance difference among the

TABLE VI

RESULTS OF TdSV USING USER-DEFINED PASSPHRASE IN TASK 2. RESULTS ARE SORTED BASED ON THE MINDCF. # SUB. MEANS NUMBER OF SUBMISSIONS.

Team IDs	# Sub.	EER [%]	MinDCF <sup>0.01<sub>norm</sub></sup>
Team 02	4	<b>1.033</b>	<b>0.0342</b>
Team 05	14	2.414	0.0670
Team 04	2	2.715	0.1424
Team 01	5	8.109	0.1482
Team 08	1	4.144	0.1589
Team 06	1	17.450	1.0000

TABLE VII

DETAILED RESULTS FOR THE BEST PERFORMING TEAM (I.E. TEAM 02) IN TASK 2.

Condition	EER	MinDCF
Overall Results	1.03	0.0342
Male	0.49	0.0275
Female	1.36	0.0376
Farsi	1.04	0.0312
English	1.05	0.0379
TC-vs-IC	1.02	0.0337
TC-vs-TW	1.08	0.0359

phrases in the dataset.

### C. Task 2 Overall Results

The overall results of the six teams participating in Task 2 are shown in Table VI. Based on the results, the first team has a significant lead. The results of the second team are somewhat acceptable based on DCF, but the rest are considerably worse. The next three teams have almost similar DCF results. Upon reviewing the detailed results of the last team (i.e., Team 05), it was found that this team failed to identify the TW tests, leading to inferior results, while they achieved suitable results for the TC-vs-IC tests.

1) *Detailed Results of Best Performing Team in Task 2:* In this section, we present the detailed results of the top-ranked team for Task 2 (i.e., Team 02), similar to Task 1. Table VII displays the detailed results, and Figure 3 illustrates the DET curve for various conditions in this task.

Based on the results in Table VII, it is clear that for this task, the results for Team 02 are much better for males than for females, although the difference is less than in Task 1. Another notable point is that the performance for both Persian and English are very close (especially based on EER), likely because the phrases in the test dataset for both languages are sufficiently long and do not negatively impact the performance.

Another important point to note from the results of this team and others for this task is the relatively low performance in rejecting TW tests. The main reason is that test phrases are not included in the training dataset, preventing teams from treating phrase recognition as a classification problem. Consequently, rejecting TW tests is significantly more challenging than in Task 1. However, the results obtained for this type of test are within the same range as those of IC, which can be considered acceptable. It is hoped that in the future, with the definition of this issue in the challenge and the presence of test datasets, better methods for identifying TW tests for user-defined passphrases will be proposed.

2) *Analysis of Results for Different Phrases in Task 2:* Similarly to Task 1, the performance for four selected phrases in the test-set of Task 2 is illustrated in Figure 4. Here, we present only the results of the best-performing team, as the other teams did not achieve

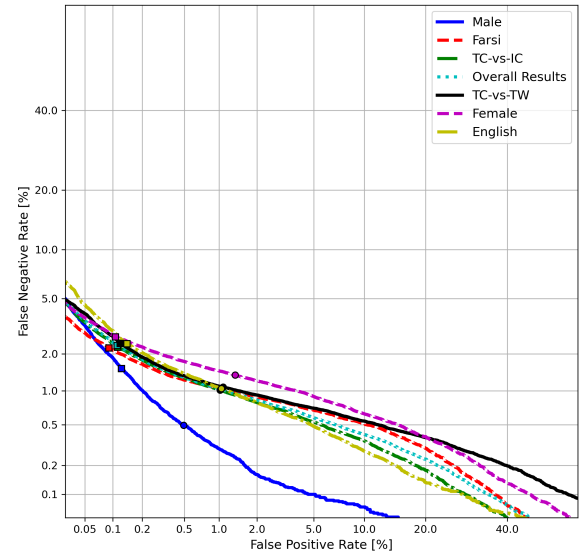


Fig. 3. DET curves of the best performing system for Task 2

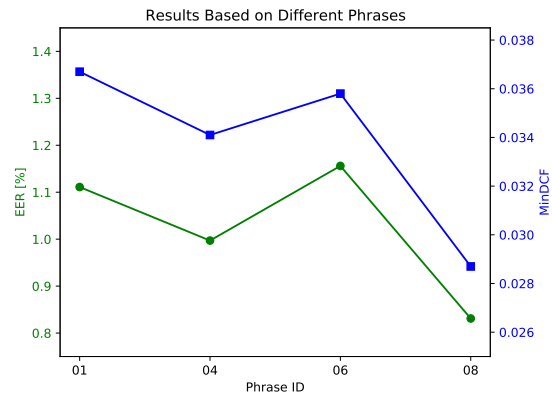


Fig. 4. Results of different phrases for the best performing team in Task 2

comparable results. A consistent trend is observed for both criteria, with most phrases exhibiting similar performance.

Comparing these plots with those in Figure 2 reveals that the performance based on phrases in this task is within the same range as in Task 1, or even better for Phrase 08. Thus, this team effectively handled user-defined passphrase scenario.

## IV. CONCLUSIONS

Text-independent speaker verification has seen significant advancements in recent years, primarily due to the availability of large datasets and the organization of various competitions in this field. Our goal with the TdSV challenge was to draw researchers' attention to the text-dependent task, fostering similar progress in this area. In this paper, we provided a comprehensive description of the two tasks included in the challenge and presented the results obtained by the participating teams. Detailed results for the best-performing team in each task were also discussed. Additionally, analysis of the phrase-specific results indicated minimal performance variation among different phrases, with most achieving comparable outcomes. Although the results from this year's challenge surpassed those of previous years, further improvements are necessary for the practical deployment of a TdSV system. We hope this challenge contributes positively to the scientific community in speaker recognition.

## REFERENCES

- [1] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: A large-scale speaker identification dataset, in: *Proc. Interspeech 2017*, 2017, pp. 2616–2620.
- [2] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: *Proc. Interspeech 2018*, 2018, pp. 1086–1090.
- [3] A. Brown, J. Huh, J. S. Chung, A. Nagrani, D. Garcia-Romero, A. Zisserman, VoxSRC 2021: The third voxceleb speaker recognition challenge, in: *arXiv preprint arXiv:2201.04583*, 2022.
- [4] J. Huh, A. Brown, J.-w. Jung, J. S. Chung, A. Nagrani, D. Garcia-Romero, A. Zisserman, VoxSRC 2022: The fourth VoxCeleb speaker recognition challenge, in: *arXiv preprint arXiv:2302.10248*, 2023.
- [5] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero, et al., The 2019 NIST speaker recognition evaluation CTS challenge., in: *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 266–272.
- [6] S. O. Sadjadi, C. Greenberg, E. Singer, L. Mason, D. Reynolds, The 2021 NIST speaker recognition evaluation, in: *arXiv preprint arXiv:2204.10242*, 2022.
- [7] H. Zeinali, K. A. Lee, J. Alam, L. Burget, SdSV challenge 2020: Large-scale evaluation of short-duration speaker verification., in: *Proc. Interspeech 2020*, 2020, pp. 731–735.
- [8] H. Zeinali, K. A. Lee, J. Alam, L. Burget, Text-dependent Speaker Verification (TdSV) Challenge 2024: Challenge evaluation plan, *arXiv preprint arXiv:2404.13428* (2024).
- [9] A. Larcher, K. A. Lee, B. Ma, H. Li, Text-dependent speaker verification: Classifiers, databases and RSR2015, *Speech Communication* 60 (2014) 56–77.
- [10] H. Zeinali, H. Sameti, T. Stafylakis, DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English., in: *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.
- [11] H. Zeinali, L. Burget, J. Cernocky, A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database, in: *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.
- [12] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, LibriSpeech: an ASR corpus based on public domain audio books, in: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [13] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.
- [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [15] J. Thienpondt, B. Desplanques, K. Demuyne, The IDLab VoxSRC-20 submission: Large margin fine-tuning and quality-aware score calibration in DNN based speaker verification, in: *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5814–5818.
- [16] Y. Zheng, Y. Zhang, C. Niu, Y. Zhan, Y. Long, D. Xu, Score calibration based on consistency measure factor for speaker verification, in: *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12371–12375.
- [17] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems* 33 (2020) 12449–12460.
- [18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., WavLM: Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing* 16 (6) (2022) 1505–1518.
- [19] B. Desplanques, J. Thienpondt, K. Demuyne, ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, in: *Proc. Interspeech 2020*, 2020, pp. 3830–3834.