



Unified Text and Speaker Verification using SSL model for Text-Dependent Speaker Verification

Nathan Griot^{1,2}, Driss Matrouf⁴, Raphael Blouet^{2,3}, Jean-François Bonastre^{1,4}, Ana Mantecon³

¹Laboratoire d'informatique d'Avignon, France

²Ardelan, France

³Daon, Ireland

⁴Defense&Security Department, Inria, France

firstname.lastname@univ-avignon.fr, firstletteroffirstnamelastname@Daon.com

Abstract

The work presented in this article falls within text-dependent speaker recognition. In our framework, each speaker owns and pronounces a secret phrase. It corresponds to two tasks: verification of the spoken text (Text Validation) and verification of the speaker's identity (SV). These tasks are usually carried-out in tandem by two different systems. Maintaining two systems involves a certain amount of complexity and may present shortcomings in terms of reliability. In this paper, we propose to use a Self-Supervised Learning Model (SSL) to develop a unified system capable of performing both tasks simultaneously. The proposed approach combines two models on a common SSL and takes advantage of a teacher-student paradigm to integrate textual constraints into the SV part, without requiring lexical labels during its learning phase. Evaluation on different datasets demonstrates the effectiveness of the approach.

Index Terms: Speaker verification, Spoken text validation, SSL model, Text-Dependent Speaker Verification

1. Introduction

Speaker Verification can be divided into two categories: Text-Independent Speaker Verification (TI-SV) and Text-Dependent Speaker Verification (TD-SV). Both approaches aim to verify whether the two given audio recordings are produced by the same speaker. However, TI-SV focuses solely on speaker verification, while TD-SV extends this by verifying the lexical content. Lexical verification offers an advantage in security systems as it can be used for two-step authentication. TD-SV task itself can be further subdivided into three distinct approaches. It can be tailored to a specific lexical context using a singular keyword with wake-up word detection such as "Okay xxx" [1], it can also be tailored to a limited number of words or pass-phrases such as voice commands "turn on the lights", and finally, it can handle any lexical content.

Several approaches have been used for the TD-SV task. One of the first solutions is to integrate phonetic information by using phoneme classification as an auxiliary task to preserve lexical information [2]. The phone classification auxiliary task is done at the frame level while the speaker is done on at the segment level. This approach performs well on a known set of pass-phrases such as voice commands but it is limited by the ability of a simple phone recognition tool to model the complex effects of coarticulation. To overcome this limitation, a potentially costly adaptation of the auxiliary phone classifier may be necessary when new commands are introduced. Another approach is to introduce the text constraints in the SV system by using **speaker + lexical** classes. Therefore a speaker can have

multiple classes depending on the lexical content pronounced [3, 4, 5]. This can be achieved directly during the training phase (or using a fine-tuning) of the system by changing the classification head or introducing a PLDA backend. This approach requires the use of substantial databases, labeled in lexical content and speaker, with strong constraints, such as the fact that each sentence must be pronounced several times by the same speaker and by other speakers. Finally, Text validation system can be used as a lexical content filtering approach. They are used to filter non-matching lexical content. Recently an ASR filtering approach has been proposed using a Fast Conformer-based Automatic Speech Recognition (ASR) model [6].

This work focuses on TD-SV where the system can handle any passphrase. In recent deep-learning SV approaches derived from xVectors [7, 8], handling the lexical content requires a large training dataset with data labeled at speaker and lexical content level. In addition, a lexical content must be pronounced several times by the same speaker, but also by a sufficient number of other speakers. To the best of our knowledge, no such data sets exist, at least on a sufficiently large scale.

To address this issue we introduce a novel approach by leveraging knowledge distillation through teacher-student approach [9]. The proposed teacher-student approach is a two-step learning process. First, the teacher model is trained for a text validation task using a large-scale dataset with lexical labels and no speaker labels. Then, this lexical knowledge is transferred to the student model. The student model is trained for a speaker recognition task using a dedicated dataset with only speaker labels. It learns the lexical knowledge directly from the teacher by using the MSE loss, comparing the teacher and student embeddings. Once the student has been trained, it can be used for both tasks, text validation and speaker recognition. This unification reduces computation costs and can improve the management and reliability for commercial applications. It can also improve performance as multitasking is known for its potential to enhance a specific task in speech processing [10].

In this work, we propose to exploit the SSL models for the teacher and the student model. SSL models have been widely used due to their training methodology, SSL models are trained on a diverse and substantial amount of data, eliminating the need for hard labels. This allows the models to adapt to a wide range of tasks. SSL models have demonstrated their versatility in various applications including natural language processing, image recognition, object classification, and more [11, 12]. Recent works have shown that such models can solve many speech processing tasks [13, 14, 15]. Particularly, it is suggested to use lower layers for speaker recognition and upper layers for lexical information [16].

This paper is organized as follows. Section 2 presents

the experimental setup, including the datasets used for training and evaluation. Next, Section 3, describes the baseline system (text validation and speaker recognition) along with the corresponding results. Subsequently, Section 4, presents the adopted methodology and the proposed system architecture. In section 5, the overall results are compared to the baseline and explained. Finally, in Section 5, we will summarize the approach, its results, and outline future directions.

2. Experimental Setup

In the context of TD-SV, the system has to accept only trials that have the same speaker and the same lexical content while refusing any other combination. This will be achieved by using two systems: **Text validation** system followed by **TI speaker verification** system. In this paper the speaker verification performance will be presented only for test pairs having the same lexical content. This section presents a description of the training and evaluation procedure of both tasks.

2.1. Dataset description

In this work three datasets are used, VoxCeleb¹, DeepMine [17], Common Voice². Voxceleb2 is a large-scale speaker recognition dataset obtained by extraction of open-source media. It has over 1 million utterances from around 6,000 speakers. This dataset is widely used for TI-SV for training while Voxceleb1 is used for evaluation.

DeepMine, is a multipurpose dataset, such as text-prompted speaker verification, TD-SV or TI-SV. It is also a multilingual dataset with both English and Persian audio recordings from over 1,400 speakers and more than 350,000 recordings. Finally, we have CommonVoice made by Mozilla, similar to Voxceleb and DeepMine it is a large-scale dataset with multilingual recordings. Common Voice recordings come with their corresponding transcription which is key for our work.

2.2. Training datasets

In this work, out of the three datasets, Voxceleb and Common Voice will be used for training. **For speaker verification** system training we use VoxCeleb2 dataset, which contains a high variety of speakers with multiple recordings per speaker. On the other hand, **for text validation**, we use Common Voice dataset, which contains broad lexical contents with multiple repetitions. We proposed a training set on Common Voice by using the transcriptions. We regrouped passphrases based on the transcription and kept only passphrases repeated at least five times and less than twenty times. This gave us close to 10,000 unique passphrases for training. Using these 10,000 passphrases we can make a classification head and extract an embedding representing the lexical content. The number of files shown in Tab. 1 is before any augmentation. Using Common Voice we are doing on-the-fly augmentation using Musan dataset [18].

Table 1: *Train corpus description. TV: Text-Validation, SV: Speaker verification*

Corpus	Purpose	nb files	utterance	spks	Dur(s)
CV EN	TV	158k	10.3k	-	3.25
VOX-2	SV	1.09M	-	6k	3

¹<https://www.robots.ox.ac.uk/vgg/data/voxceleb/>

²<https://commonvoice.mozilla.org/>

2.3. Evaluation datasets

The evaluation will be conducted on two datasets, DeepMine and Common Voice (description in Table 2). Common Voice will be used to assess the robustness of the text validation system in both English and French evaluation sets. The DeepMine dataset will be used to evaluate both the text validation system (in English and Persian) and the speaker verification system. In this paper, all experiments regarding speaker verification are performed on recording pairs in which the lexical content is the same. Our approach is based on the fact that the text validation system filters out pairs of recordings with different lexical contents. For this reason, the DeepMine speaker evaluation dataset contains exclusively pairs with identical lexical contexts.

Table 2: *Evaluation corpus description. TV: Text-Validation TD-SV: Text-dependent Speaker Verification.*

Corpus	Purpose	Trials	utterance	Dur (s)
CV EN	TV	268k	1740	3.25
CV FR	TV	381k	225	3.47
DM EN	TV & SV	27k	5	2.30
DM EN	TD-SV	27k	5	2.30
DM FA	TV	25.4k	5	1.9
VOX-1	SV	37k	/	-

All evaluations use the Equal-Error Rate metrics (EER) and the Tandem Equal-Error Rate (T-EER) for the TD-SV trial [19].

3. Baseline

3.1. Text Validation

The Text validation baseline system uses the ResNet34 architecture, where the target class is the sentence ID. The system is trained on a subset of English Common Voice described in 2.2. All system parameters are optimized using the Additive Angular Margin Loss (AAM) [20]. The amount of available data and unique passphrases makes it possible to train the text validation system from scratch. This system generates an embedding that represents the lexical content. For evaluation, we use the cosine similarity backend process using the embeddings to generate a score used to accept or reject the pair of utterances. Table 3 shows the performance in the Common Voice and DeepMine evaluation datasets in two languages (English and French). The performances are very good in all cases, especially for English which is the language used during training.

Table 3: *EER of the text validation task system evaluation performances on DeepMine and Common Voice for the lexical verification task.*

	DM EN	DM FA	CV EN	CV FR
ResNet34	0.21%	2.2%	0.28%	2.4%

3.2. TI Speaker verification

Our baseline TI-SV will be the ReDimNet architecture [21]. We chose the ReDimNet because of its recent introduction and demonstration of state-of-the-art results for TI-SV. Additionally, we used the provided ReDimNet, ensuring that all of the following results using the ReDimNet can be replicated³.

³<https://github.com/IDRnD/ReDimNet>

Table 4: EER of the ReDimNet system evaluation performances on DeepMine and VoxCeleb1 for the TI-SV task.

	Voxceleb1	DM EN
ReDimNet	0.49%	4.55%

4. Unified TD-SV SSL

For our SSL choice, we decided to use the large WavLM SSL model [16]⁴. The WavLM model is an SSL model used for speech processing. It was successfully applied on multiple speech processing tasks such as speaker verification and semantics. For TI speaker verification, the SSL models have been used in combination with Multi-Head Factorized Attention pooling (MHFA) backend [22]. In this work, we will apply the MHFA for the speaker verification and for the text validation backend. For training, the VoxCeleb2 dataset will be used as main training source, removing the need for transcription or lexical identification. To do so we will utilize the teacher-student learning approach. A teacher system is trained on Common Voice for text validation, the teacher generates lexical embedding used as soft labels for the student system.

4.1. Text validation Teacher SSL

The teacher architecture can be seen in Fig.1 and is structured as follows: the core component is the large WavLM model, which consists of 24 layers. The outputs from all 24 layers are extracted and fed into the MHFA. The embedding produced by the MHFA represents the lexical content, it is then provided to a classification head using AAM loss. The text validation embedding is also necessary in evaluation using cosine similarity. Table 5 shows the text validation results of the teacher system on all evaluation datasets. Comparing Table 5 and Table 3, we can see a clear superiority of SSL over ResNet for text validation, especially for English which is the language used for training.

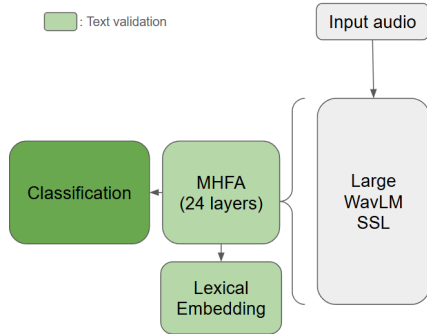


Figure 1: Teacher SSL architecture. The teacher SSL is fine-tuned only for the text validation task.

Table 5: EER of the Text validation task system evaluation performances on DeepMine and Common Voice for the lexical verification task.

	DM EN	DM FA	CV EN	CV FR
SSL teacher	0.07%	2%	0.17%	2.9%

⁴<https://huggingface.co/microsoft/wavlm-large>

4.2. Multitask Student SSL

The student SSL model is a multitask system performing both speaker verification and text validation. During training, two learning rates are employed: one for fine-tuning the WavLM model, set at 10^{-6} , and another for the rest of the system, set at 10^{-3} . Both tasks use the WavLM Large model as a common block.

In Fig. 2 we can see, firstly, the speaker block in blue, it uses layers 7 to 12 of the WavLM system for the MHFA block. The MHFA block compresses the input into an embedding, which serves as the speaker representation during evaluation. During training, this embedding is fed into a classification head with 5994 classes using AAM loss, corresponding to the number of unique speakers in the VoxCeleb2 dataset.

Secondly, we have the text validation block in green (see Fig. 2). The text validation block employs an MHFA block similar to the one used in the speaker block; however, unlike the speaker block, this MHFA utilizes layers 1 to 24 of the large WavLM model. The MHFA reduces this information to an embedding. The resulting lexical embedding is compared to the lexical embedding from the teacher using Mean Squared Error (MSE) loss. The text validation block is trained by using the MSE criterion between the teacher-given lexical embedding and the student-given lexical embedding. The entire system is then trained by optimizing the sum of the two objective functions: MSE loss for the lexical part and AAM loss for the speaker part.

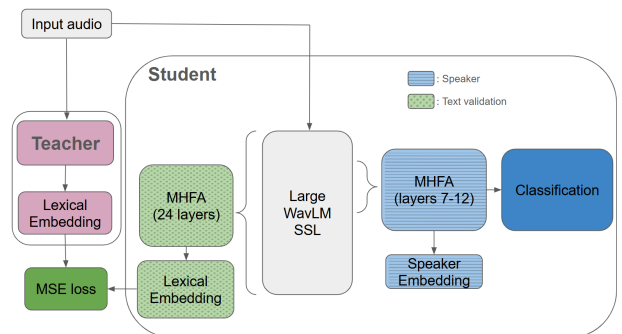


Figure 2: Full architecture, presenting both student and teacher architecture. The student is trained on VoxCeleb2 with blue side for speaker classification and green side using MSE loss on lexical embeddings.

5. SSL Student performances

5.1. Text validation

All text validation results can be found in Table 6. From these results, we can see that the student outperforms on all evaluation sets both the ResNet34 baseline but also the teacher SSL. The teacher-student approach showed its effectiveness by having the student outperforming the teacher. In this work the teacher-student approach was used to remove the need for hard-labeled data on lexical content. This approach seems effective, but the exact reason why it was so effective needs to be further tested, however, we think it is due to the soft labels being richer and reducing the risk of overfitting. Also, we believe the multitask approach helped preserve information resulting in a more efficient lexical embedding.

Table 6: Text verification EER. Comparing ResNet34 architecture and SSL student and teacher text verification. Performances are evaluated in three different languages, English, French, and Persian.

Model/EER%	Common Voice EN	Common Voice FR	DeepMine EN	DeepMine FA
ResNet 34	0.28	2.4	0.21	2.43
SSL teacher	0.17	2.9	0.07	2
SSL student	0.1	1.7	0.03	1.4

5.2. Speaker verification

All speaker verification results can be found in Table 7. The SSL student outperforms the ReDimNet on the DeepMine lexical subset. On the other side the ReDimNet outperforms on VoxCeleb1. In our use case the DeepMine dataset is more representative as it contains only matching lexical pairs. As previously stated the text validation is used as a filter, filtering out all non-matching lexical content pairs. Evaluating the speaker system on lexical matching pairs is more representative of the real use case.

We believe that performing multitask training helped the speaker verification system when the lexical content is the same. As we force the lexical information to be retained in the SSL layers during finetuning for the text validation task. This information is likely used for speaker verification gaining performance when the lexical content is identical. Furthermore, this may also explain the performance degradation in the case of VoxCeleb1 where the lexical content does not match in the test pairs recordings.

Table 7: EER of the ReDimNet and SSL student system evaluation performances on DeepMine and VoxCeleb1 for the TI-SV task

	VoxCeleb1	DM EN
ReDimNet	0.49%	4.55%
SSL student	1.29%	3.49%

5.3. Text-dependent speaker verification

For TD-SV evaluation, we will use tandem-EER (T-EER) [19]. T-EER is a variation of the well-known EER, it is used for tandem architectures. In our case, text validation is followed by speaker verification. The following table 8 and figure 3,4 demonstrate the performance on TD-SV evaluation set using this metric. In figures 3,4 the intersection of the blue line represented with a pink dot is the optimal threshold for the tandem system. The T-EER is then computed using those thresholds. Table 8 resume the results given by the figures. We can notice first that the separation between matching and non-matching lexical content is far better for the SSL and secondly that the results are similar to the TI-SV results. This is an effect of the speaker verification system being the most limiting system between the two.

Table 8: T-EER on the baseline against the unified SSL on DeepMine English TD-SV evaluation set.

System/Concurrent T-eer	DM EN
baseline (ReDimNet + ResNet34)	4.28%
SSL student	3.46%

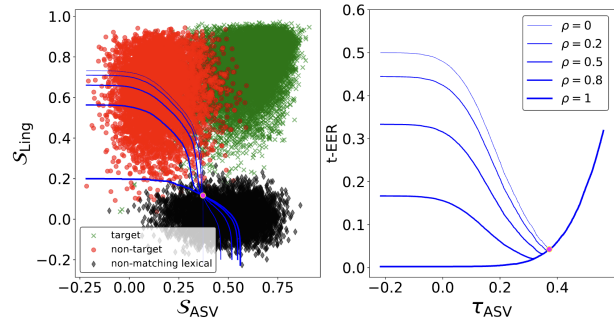


Figure 3: T-EER evaluation using both baseline models Text validation: ResNet34, Speaker verification: ReDimNet. Y-axis represent the lexical verification score, X-axis represents the Speaker verification score

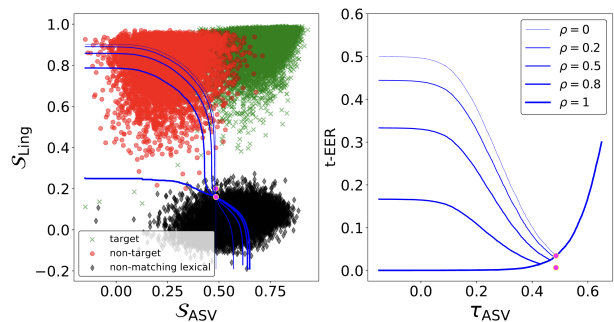


Figure 4: T-EER evaluation using both unified SSL TDSV sub-tasks. Y axis represents the lexical verification score, X-axis represents the Speaker verification score

6. Conclusion

In this article, we proposed to use SSL models to develop an unified system capable of doing both text validation and speaker verification tasks. We implemented teacher-student approach to bypass data limitation allowing us to leverage lexical information from unlabeled speech data. This approach makes it easy and cheap to adapt to new languages. We have demonstrated the effectiveness of the proposed approach by introducing an experimental protocol using several databases and several languages. We have shown that the student outperforms both the challenging baseline and the teacher system for text validation. At the same time, the student system outperforms the baseline speaker recognition system on DeepMine evaluation set. The proposed approach proposes a tandem verification using two tasks (namely speaker and text verification) of the unified system. We believe that the interaction between these two tasks can be mutually beneficial, enhancing the overall performance. For this reason, future work will focus on further exploring and optimizing the interaction between the two tasks.

7. References

- [1] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5115–5119.
- [2] Y. Liu, Z. Li, L. Li, and Q. Hong, "Phoneme-Aware and Channel-Wise Attentive Learning for Text Dependent Speaker Verification," in *Proc. Interspeech 2021*, 2021, pp. 101–105.
- [3] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," in *Interspeech 2013*, vol. 500, 08 2013.
- [4] Z. Chen and Y. Lin, "Improving x-vector and plda for text-dependent speaker verification," in *INTERSPEECH*, 2020, pp. 726–730.
- [5] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Interspeech 2015*, 2015, pp. 185–189.
- [6] M. Molavi and R. Khodadadi, "The svasr system for text-dependent speaker verification (tdsv) aac challenge 2024," 2024. [Online]. Available: <https://arxiv.org/abs/2411.16276>
- [7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. 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.
- [8] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] C. Hu, X. Li, D. Liu, H. Wu, X. Chen, J. Wang, and X. Liu, "Teacher-student architecture for knowledge distillation: A survey," 2023. [Online]. Available: <https://arxiv.org/abs/2308.04268>
- [10] Y.-C. Chen, S. wen Yang, C.-K. Lee, S. See, and H. yi Lee, "Speech representation learning through self-supervised pre-training and multi-task finetuning," 2021. [Online]. Available: <https://arxiv.org/abs/2110.09930>
- [11] A. Ciocarlan, S. Lefebvre, S. L. Hégarat-Masclé, and A. Woiselle, "Self-supervised learning for real-world object detection: a survey," 2024. [Online]. Available: <https://arxiv.org/abs/2410.07442>
- [12] L. C. Huang, D. J. Chiu, and M. Mehta, "Self-supervised learning featuring small-scale image dataset for treatable retinal diseases classification," 2024. [Online]. Available: <https://arxiv.org/abs/2404.10166>
- [13] C. Wang, Y. Wu, S. Chen, S. Liu, J. Li, Y. Qian, and Z. Yang, "Self-supervised learning for speech recognition with intermediate layer supervision," 2021. [Online]. Available: <https://arxiv.org/abs/2112.08778>
- [14] Z. Chen, N. Kanda, J. Wu, Y. Wu, X. Wang, T. Yoshioka, J. Li, S. Sivasankaran, and S. E. Eskimez, "Speech separation with large-scale self-supervised learning," in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [15] D. Combei, A. Stan, D. Oneata, and H. Cucu, "Wavlm model ensemble for audio deepfake detection," 2024. [Online]. Available: <https://arxiv.org/abs/2408.07414>
- [16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022. [Online]. Available: <http://dx.doi.org/10.1109/JSTSP.2022.3188113>
- [17] H. Zeinali, H. Sameti, and T. Stafylakis, "Deepmine speech processing database: Text-dependent and independent speaker verification and speech recognition in persian and english," in *The Speaker and Language Recognition Workshop*, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:51743953>
- [18] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.
- [19] T. Kinnunen, K. Lee, H. Tak, N. Evans, and A. Nautsch, "t-eer: Parameter-free tandem evaluation of countermeasures and biometric comparators (to appear)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [20] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 5962–5979, Oct. 2022. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2021.3087709>
- [21] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape dimensions network for speaker recognition," in *Interspeech 2024*. ISCA, Sep. 2024, p. 3235–3239. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2024-2116>
- [22] J. Peng, O. Plchot, T. Stafylakis, L. Mosner, L. Burget, and J. Cernocky, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," 2022. [Online]. Available: <https://arxiv.org/abs/2210.01273>