INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Conformer-based Language Embedding with Self-Knowledge Distillation for Spoken Language Identification

Feng Wang[1], *Lingyan Huang*[1], *Tao Li*[1],*Qingyang Hong*[1*], *Lin Li*[2*]

[1]School of Informatics, Xiamen University, China
[2]School of Electronic Science and Engineering, Xiamen University, China
{qyhong, lilin}@xmu.edu.cn

Abstract

The utilization of Conformer-based architecture has been shown to be effective in improving the performance of spoken language identification (LID) in recent years due to Conformer's superior representational capacity. However, when performing language identification on short speech segments, a significant drop in performance is often observed. In this paper, we adopt a method to alleviate this issue by introducing a self-knowledge distillation technique to Conformer-based LID architecture whose encoder was pretrained by an ASR task. We distill the predictive distribution between the original input and the input processed by a double-ended random masking module during the training stage for each sample. Experimental results demonstrate the effectiveness of the method on two datasets: OLR21 with 16,000 Hz sampling rate and LRE22 with 8,000 Hz sampling rate. Moreover, the method also enhances the performance of language identification on short-duration speech segments.

Index Terms: language identification, short utterances, self-knowledge distillation

1. Introduction

Spoken language identification (LID) is the task of determining the language of an utterance. And it is an essential task in speech processing, especially in multi-lingual applications. Reliable and robust LID is critical for achieving high performance and accuracy in these systems [1, 2].

The LID system is typically composed of two primary components: an language embedding extractor and a back-end scoring method. The embedding extractor maps variable-duration speech utterances to fixed-dimensional language representations that capture high-dimensional language features. Subsequently, the back-end scoring method measures the similarity of the language representations to identify the language of the utterance. Over the years, various neural network architectures have been employed as embedding extractors in LID systems, including E-TDNN [3] , ResNet [4], and ECAPA-TDNN [5]. These architectures have demonstrated success in extracting meaningful features from speech data and providing high-quality language representations.

Recent studies have proposed that there have been significant advances in neural network architectures for LID tasks [6, 7, 8, 9, 10]. Conformer-based [11, 12] models gain popularity due to their ability to model both local and global dependencies in an audio sequence in a parameter-efficient way [13, 14, 15]. Moreover, using a pretrained automatic speech recognition model to provide informative speech representations has proven to be effective for the downstream LID task. It's worth noting that reference [16] utilizes a transfer learning approach, where a U2++ [5] encoder-decoder model is pretrained and then the encoder is further fine-tuned for the LID task. By transferring knowledge from related speech tasks or high-resource languages, transfer learning can boost the performance of LID systems. Furthermore, many studies have explored the use of different loss functions and regularization techniques to train embedding extractors that can generalize well to unseen data [17].

In recent years, many methods have been proposed to learn more efficient representation. Knowledge distillation from pretrained deep networks suggests that we can use more information from the soft target probability to train other neural networks or train them based on the soft target probabilities of the training model itself [18]. Self-knowledge distillation has been shown to be effective in both Natural Language Processing (NLP) and Computer Vision (CV) areas [18, 19, 20, 21, 22]. Self-distilled self-supervised speaker representation learning has pushed the performance of the speaker representation to a new limit[23]. R-Drop [21] forces the output distributions of different submodels generated by dropout to be consistent with each other and is universally effective in several tasks, including neural machine translation, language understanding, language modeling, and image classification. A dual-mode framework with knowledge distillation was proposed to enhance the LID performance on various-duration speech [24].

Inspired by these works, we trained a multilingual ASR model and employ the trained Conformer encoder to the LID model and adopt a segment mask self-knowledge distillation (SM-KD). For each sample, we distill the predictive distribution between the origin audio segment and truncated audio during training. Consequently, it not only improves the model's generalization ability but also facilitates the learning of feature extraction for short-duration speech. We evaluated the method on the OLR21 [25] dataset with 16,000 Hz sampling rate and the LRE22 [26] dataset with 8,000 Hz sampling rate. The experimental results showed that this method could improve LID performance, especially in short audios.

This paper is organized as follows: Section 2 outlines the ASR pretrain and SM-KD method to the LID task. In Section 3, we provide a comprehensive description of the experimental setup utilized for training and evaluation in this study. Section 4 presents an analysis of the experimental results, while Section 5 summarizes the method and its efficacy.

2. Methods

The LID system employed in this paper consists of two primary components: a front-end feature extraction module and a back-

*Corresponding author

10.21437/Interspeech.2023-1557

end classifier. This front-end and back-end framework has been widely used in both LID and speaker recognition tasks. In this section, we describe the Conformer-based LID front-end model. Additionally, we introduce a segment mask self-knowledge distillation method, which leverages a form of knowledge distillation to enhance the performance of the front-end model. Finally, we present the back-end classifier used in our system for language recognition.

2.1. Conformer-based LID model

The backbone of the LID Conformer system is shown in Figure 1. With the Conformer encoder consisting of subsampling layers, a stack of Conformer blocks models the frame-level language representation. An attentive statistics pooling (ASP) [27] processes all the information across the time dimension and maps frame-level representations into segment-level features, which are then projected by a linear layer with batch normalization and nonlinear activation to the utterance level embedding. This structure is trained to classify these languages on the training data with a cross-entropy (CE) loss. The Conformer block is composed of two feed-forward modules (FNN), similar in structure to the Macaron block, with residual connections. Multi-head self-attention (MHSA) and Convolution modules are sandwiched between the two FNN modules. MHSA captures the global information of the input sequence, while Convolution provides the network with an inductive bias. This combination of modules allows the Conformer block to learn both local and global representations of the input, making it a powerful building block for various tasks in speech processing, including language identification.

Attentive statistics pooling is a technique used in deep learning for extracting features from sequential data, such as speech or text. It involves computing a weighted average of the input sequence, where the weights are learned through an attention mechanism. The attention mechanism allows the model to selectively focus on different parts of the input sequence, giving more weight to the most relevant parts for the task at hand. Another advantage of using the Conformer architecture is that the main component of the model lies in its widespread application across various audio tasks [11, 14, 28, 29], enabling convenient knowledge transfer for enhancing model capacity.

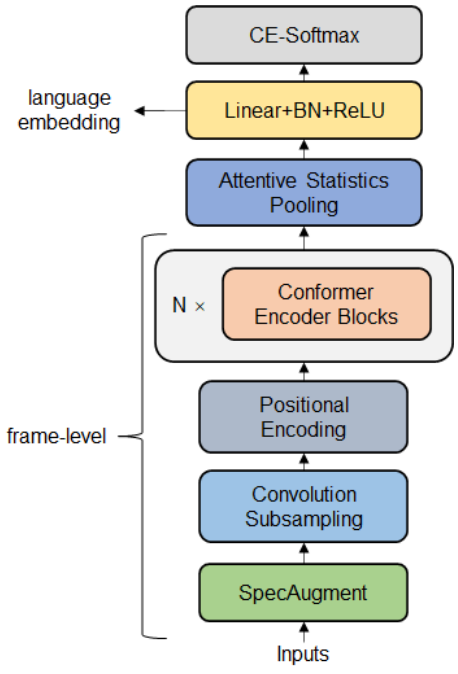

Figure 1: *Schematic diagram of Conformer-based LID backbone.*

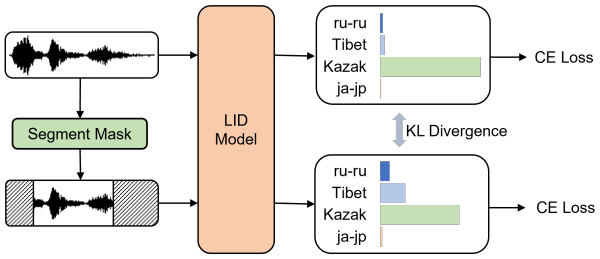

Figure 2: *Illustration of segment mask self-knowledge distillation architecture. The input will go through the model twice and obtain two distributions P and Q. In the second pass, the input is masked to a shorter audio segment.*

2.2. Segment mask self-knowledge distillation

The Conformer-based LID model exhibits a superior representational capacity that facilitates easy memorization of the training set. However, this attribute also causes overfitting problems when the training data is inadequate. In the training stage, determining the chunk size of input samples is crucial. When the chunk size is set too large, the model's ability to extract features from short speech segments is significantly diminished. Conversely, if the chunk size is set too small, the amount of information provided by each sample will be insufficient, thus posing difficulty in training and convergence of the model.

To mitigate this problem, we employ a simple approach that leverages self-knowledge distillation for LID tasks. We introduce the segment mask self-knowledge distillation method to Conformer-based model to enhance the performance of the LID system.

As shown in Figure 2, during the forward propagation stage, the input undergoes two passes through the model. The first pass produces a probability distribution denoted as P after processing the input x. Subsequently, a random length mask is applied to both ends of the same sample x, and the remaining segments x' can be considered as contiguous audio segments extracted from the original sample. Another distribution Q can be obtained after feeding the segmented samples into the network. To encourage the learned features to be as similar as possible, the Kullback-Leibler divergence D_{KL} is computed to measure the similarity between these two vectors:

$$D_{KL}\left(P||Q\right) = \sum P(i) \log\left(P(i)/Q(i)\right) \tag{1}$$

$$\mathcal{L}_{KL} = D_{KL}\left(P||Q\right) + D_{KL}\left(Q||P\right) \tag{2}$$

After obtaining the probability distributions P and Q, we compute the cross-entropy loss for each distribution, respectively, and calculate a weighted sum of the two CE losses as $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = \mathcal{L}_{ce}\left(P, T\right) + \mathcal{L}_{ce}\left(Q, T\right) \tag{3}$$

Algorithm 1 segment mask self-knowledge distillation

Initialize parameters θ.
while θ has not converged **do**
Sample a batch (x, y) from the training dataset.
Mask x from the left and right ends with random length to get (x', y)
Feed x and x' into the language embedding model to obtain the distributions P and Q.
Update parameters θ by computing the gradients of the loss function in Equation (4).
end while

where T denotes the ground truth. Incorporating the segment mask self-knowledge distillation strategy, the modified loss function $\mathcal{L}_{SM\text{-}KD}$ is defined as follows:

$$\mathcal{L}_{SM\text{-}KD} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} \tag{4}$$

The second input can be viewed as a truncated version of the original input and is a short-duration audio. In this structure, the first input can be considered as the teacher model to assist in the training of the truncated audio, resulting in an embedding extractor that is robust to short-duration speech. The full training procedure with the loss is summarized in Algorithm 1.

2.3. Scoring methods

The embeddings are computed from original training data and augmented audio samples. No linear discriminative analysis (LDA) method is performed. The embedding features are averaged into one enrollment embedding vector for each language. Then the logistic regression (LR) was used to compute the score of a trial on a particular language.

3. Experiments setups

3.1. Datasets

The OLR21 [25] database covers 17 languages. The training set is up to 280 hours. The sampling rate of both the training set and the test set audio is 16,000 Hz. 13 target languages (i.e., Indonesian, Japanese, Russian, Korean, Vietnamese, Mandarin, Cantonese, Sichuanese, Shanghainese, Hokkien, Tibetan, Kazakh and Uyghur) are considered in our experiment.

As illustrated in Figure 3, the test set comprises a total of 32,863 speech samples of varying lengths ranging from 1s to 33s. According to statistics, there are 3,154 samples less than 3s, 19,882 samples between 3s and 6s, and 9,827 samples longer than 6s. In order to evaluate the performance of the model on different length segments, we divided the test set into three subsets based on their duration: short, normal, and long.

In LRE 2022, the following datasets are provided by the organizer as described in [26] : 2017 NIST LRE Development Set and previous NIST LRE training data (LDC2022E16), 2017 NIST LRE Test Set (LDC2022E17), and 2022 NIST LRE Development Set (LDC2022E14), including 14 target languages and 14 non-target languages. Audios are 8-bit a-law SPHERE files sampled at 8,000 Hz. The VoxLingua107 data set [30] was also permitted for use. Only the above-specified data sets were used during training.

To enhance the robustness, the following data augmentation approaches have been adapted to improve model performance: speed perturbation, with the speed factors of 0.9, 1.0,1.1; additive noise from the MUSAN [31] dataset is mixed with the original signal, only publicly available non-speech audio and data are used; reverberation injection, using simulated room impulse responses from RIRs [32].

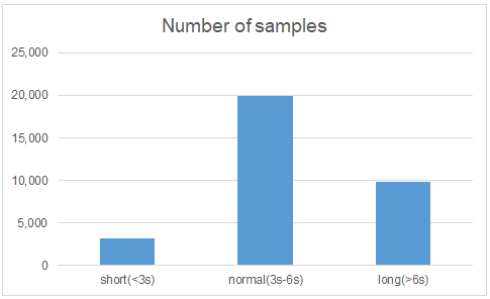

Figure 3: *Duration distribution of the OLR21 test set and partition*

3.2. Experimental settings

The ASV-Subtools platform [29, 33] was employed for training each language identification model. Feature engineering and backend processing were performed using the Kaldi platform [34], while the backbone model was trained using the PyTorch-based ASV-Subtools.

For feature engineering, 80-dimensional FBank with a frame length of 25ms and a hop size of 10ms was used. In addition, we applied cepstral mean normalization (CMN) within a 3-second sliding window. It's worth noting that no voice activity detection (VAD) was employed in the feature extraction process.

For the Conformer configures, the 12-block conformer encoder output dimension is 256, the feed-forward dimension is set to 1024, and the number of attention heads is 4. LID models are trained with ralamb optimizer. The hyperparameters used for training the model included an initial learning rate of 0.0005 and weight decay set at 1e-1. Lookahead technique was not employed during the training process. During the SM-KD training process, we set α in Equation (4) to 0.35.

3.3. Model evaluation

During the testing phase, embeddings are computed for each segment of audio through the forward pass of the front-end neural network. For back-end, we choose LR to compute the score of extracted embeddings. Evaluation performance is measured by Equal Error Rate (EER) and Cavg in the OLR 2021 datasets [25]. And for LRE22 datasets, actCprimary and minCprimary are uesd as performance measurement metric [26].

For comparison, we trained the baseline E-TDNN x-vector model and Conformer-based LID model. Their performances are reported in Table 1. As indicated here, the Conformer-based model is competitive in language recognition tasks. The conformer block is adept at capturing global information through attention mechanisms while also benefiting from the convolution-based modeling of local invariance.

4. Results and analysis

As is shown in Table 1, with the self-knowledge distillation technique, the performance of Conformer model is further improved. We partitioned the original test set into three subsets

Table 1: *Performance of Baseline, Conformer with and without ASR pretraining, and the SM-KD method on OLR2021 dataset*

Model	test		short		normal		long	
	Cavg	**EER%**	**Cavg**	**EER%**	**Cavg**	**EER%**	**Cavg**	**EER%**
Baseline	0.0817	8.977	-	-	-	-	-	-
Conformer	0.0229	2.173	0.0289	3.9	0.0219	2.055	0.014	1.74
+ SM-KD	0.014	1.43	0.0208	2.949	0.0143	1.257	0.0108	1.257
+ ASR pretrain	0.0122	1.263	0.0201	2.917	0.0123	0.0123	0.009	1.079
+ ASR pretrain + SM-KD	**0.0112**	**1.187**	**0.018**	**2.695**	**0.0117**	**1.056**	**0.0074**	**0.9667**

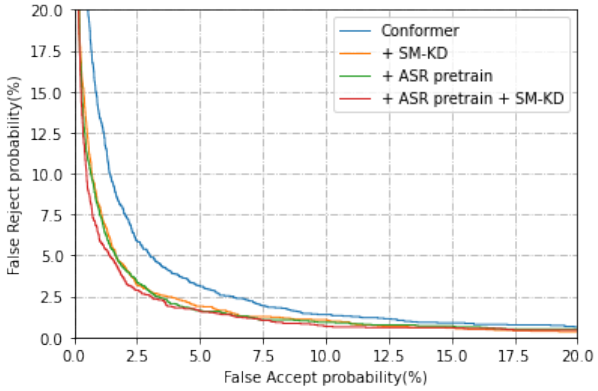

Figure 4: *DET curves of each system on short test set*

based on the length of the audio recordings. Then, we evaluated the performance of each model on these three subsets separately. Figure 4 illustrates the Detection Error Tradeoff (DET) curves of each system on the short test set. Similar to calculating EER, for each sample x and each language y, we obtain a score $score_{x,y}$. By setting a threshold θ, if the $score_{x,y}>\theta$ and the ground truth $LID(x) \neq y$, then the False Acceptance (FA) count increases. Conversely, if the $score_{x,y}<\theta$ and $LID(x) = y$, then the False Rejection (FR) count increases. We adjust the threshold θ to calculate the FAR and FRR at each threshold and plot them to obtain the DET curve. The experimental results reveal that our method can effectively enhance the performance of the LID model, particularly in short speech scenarios. The Conformer model utilizing the segment mask self-knowledge distillation outperformed the baseline Conformer model not only on the short-duration test subset but also on the long-duration test subset, demonstrating that the inclusion of the self-knowledge distillation strategy can extract more robust language embeddings. The speech sample in the OLR21 dataset contains not only the language id but also the corresponding transcript. So we trained a multilingual ASR model using the data among all target languages as the ASR-pretrained model. The experimental results also indicate that the ASR-pretrained encoder with fine-tuning has a significant performance improvement, which suggests that the encoder trained with ASR tasks has the ability to extract linguistic features and can be beneficial to LID tasks. This structure can also be applied to the fine-tuning process after ASR transfer learning, resulting in better performance of the final model.

Table 2: *The final rankings for the OLR21 constrained LID task, Only the specified data can be used in the training process*

Ranking	**Team Name**	**Cavg**	**EER%**
1	X-Voice	0.0025	0.2708
2	TalTech	0.0079	0.8642
3	funspeech	0.0083	0.9311
	Our single system	0.0112	1.187
4	Anonymous	0.0114	1.184

Table 3: *Experimental performance of Baseline, Conformer and Conformer with SM-KD on LRE22 dataset*

Model	**actCprimary**	**minCprimary**	**EER%**
ECAPA-TDNN	0.40335	0.40679	15.48
Conformer	0.27252	0.29645	9.106
+ SM-KD	**0.25104**	**0.27553**	**8.407**

Table 2 shows the top four systems and their performance in the OLR21 competition. In the Constrained LID track, Only the specified data can be used in the training process. Our single system achieved the performance of the top four systems after model fusion in the competition.

Table 3 presents the performance of ECAPA-TDNN, Conformer, and Conformer with SM-KD on LRE22 datasets. It is worth mentioning that we use a two-stage training strategy, using all data for pre-training, and fine-tuning in the second stage using the target language. We found that Conformer-based language embedding with SM-KD can also achieve better performance on the 8000 Hz sampling rate data set.

5. Conclusions

This paper introduces a segment mask self-knowledge distillation approach for the LID task based on pretrained Conformer structure. We distill the predictive distribution between the original input and the input processed by a double-ended random masking module during the training stage. We further demonstrated that the forward propagation of the original audio during training is regarded as a teacher that helps the model learn from short speech segments. Moreover, this approach improves the model's feature extraction ability and leads to more generalized embeddings. The effectiveness of the method was evaluated on OLR21 and LRE22 datasets. The experimental results demonstrated that this method improved the performance of LID, particularly in short-duration audio scenarios.

6. Acknowledgements

Thanks to the National Natural Science Foundation of China (Grant No.62276220, No.62001405 and No.61876160) for funding.

7. References

[1] K. Kukk and T. Alumäe, "Improving language identification of accented speech," *arXiv preprint arXiv:2203.16972*, 2022.

[2] P. Shen, X. Lu, and H. Kawai, "Transducer-based language embedding for spoken language identification," *arXiv preprint arXiv:2204.03888*, 2022.

[3] Y. Liu, T. Liang, C. Xu, X. Zhang, X. Chen, W.-Q. Zhang, L. He, R. Li, Y. Wu, P. Ouyang *et al.*, "Thuee system description for nist 2019 sre cts challenge," *arXiv preprint arXiv:1912.11585*, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[6] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[7] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, "On the use of deep feedforward neural networks for automatic language identification," *Computer Speech & Language*, vol. 40, pp. 46–59, 2016.

[8] P. Shen, X. Lu, L. Liu, and H. Kawai, "Local fisher discriminant analysis for spoken language identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5825–5829.

[9] P. Shen, X. Lu, S. Li, and H. Kawai, "Feature representation of short utterances based on knowledge distillation for spoken language identification." in *Interspeech*, 2018, pp. 1813–1817.

[10] A. Lozano-Diez, R. Zazo Candil, J. González Domínguez, D. T. Toledano, J. Gonzalez-Rodriguez *et al.*, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech and Communication Association, 2015.

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[12] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[13] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.

[14] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5749–5753.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] A. Lyu, Z. Wang, and H. Zhu, "Ant multilingual recognition system for olr 2021 challenge," *Proc. Interspeech 2022*, pp. 3684–3688, 2022.

[17] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[18] S. Hahn and H. Choi, "Self-knowledge distillation in natural language processing," *arXiv preprint arXiv:1908.01851*, 2019.

[19] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 876–13 885.

[20] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu *et al.*, "R-drop: Regularized dropout for neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 890–10 905, 2021.

[21] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6567–6576.

[22] D.-Q. Vu, N. Le, and J.-C. Wang, "Teaching yourself: A self-knowledge distillation approach to action recognition," *IEEE Access*, vol. 9, pp. 105 711–105 723, 2021.

[23] Z. Chen, Y. Qian, B. Han, Y. Qian, and M. Zeng, "A comprehensive study on self-supervised distillation for speaker representation learning," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 599–604.

[24] H. Liu, L. P. García-Perera, A. W. Khong, J. Dauwels, S. J. Styles, and S. Khudanpur, "Enhancing language identification using dual-mode model with knowledge distillation." in *Odyssey*, 2022, pp. 248–254.

[25] B. Wang, W. Hu, J. Li, Y. Zhi, Z. Li, Q. Hong, L. Li, D. Wang, L. Song, and C. Yang, "Olr 2021 challenge: Datasets, rules and baselines," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1097–1103.

[26] Y. Lee, C. Greenberg, E. Godard, A. A. Butt, E. Singer, T. Nguyen, L. Mason, and D. Reynolds, "The 2022 nist language recognition evaluation," *arXiv preprint arXiv:2302.14624*, 2023.

[27] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Proc. Interspeech 2018*, pp. 2252–2256, 2018.

[28] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "Df-conformer: Integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 161–165.

[29] D. Liao, T. Jiang, F. Wang, L. Li, and Q. Hong, "Towards a unified conformer structure: from asr to asv task," *arXiv preprint arXiv:2211.07201*, 2022.

[30] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[33] F. Tong, M. Zhao, J. Zhou, H. Lu, Z. Li, L. Li, and Q. Hong, "Asv-subtools: Open source toolkit for automatic speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6184–6188.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.