Interspeech 2024
1-5 September 2024, Kos, Greece

Self-supervised speaker verification with relational mask prediction

Ju-ho Kim[*1], *Hee-Soo Heo*[*2], *Bong-Jin Lee*[2], *Youngki Kwon*[2], *Minjae Lee*[2], *Ha-Jin Yu*[1]

[1]University of Seoul, Republic of Korea
[2]NAVER Cloud Corporation, Republic of Korea

wngh1187@naver.com, heesoo.heo@navercorp.com, hjyu@uos.ac.kr

Abstract

Recently, self-supervised learning (SSL) has emerged as a promising strategy for constructing speaker verification (SV) systems, effectively mitigating the cost and privacy issues associated with the labeling process. The majority of SSL-based SV systems tend to focus on utterance-level features, potentially overlooking the inherent inter-frame structure of speech. To bridge this gap, we propose the relational mask prediction (RMP), a novel loss function that encourages models to understand the relationships between frames. Additionally, we introduce a block aggregation Transformer (BA-Transformer) to enrich frame-level features. Models were trained without labels using the VoxCeleb2 development set and comprehensively evaluated using various test sets. Experimental results demonstrate that the proposed framework outperforms recent SSL-based SV systems, achieving an average performance improvement of 22.39% over the baseline across the entire evaluation dataset.

Index Terms: speaker verification, self-supervised learning, relational mask prediction, block aggregation Transformer

1. Introduction

The advent and subsequent application of deep learning methodologies have led to significant progress in the domain of speaker verification (SV) [1–3]. Supervised learning is a widely adopted approach to build SV systems. In this approach, a model is trained on labeled data to represent the identity of an utterance in a discriminative manner [1,2,4]. The success of supervised learning depends on the quality of the data labels; however, generating high-quality labels can be labor-intensive and costly. To mitigate these challenges, self-supervised learning (SSL) has emerged as a promising alternative [5–7]. The SSL approach allows models to extract meaningful information directly from data, eliminating the need for labeling processes.

In the field of SSL for SV, contrastive learning is a prominent method that learns speaker embeddings directly [8–11]. The fundamental principle of this method is to increase the similarity between positive pairs in latent space, which typically consist of different augmentations of the same utterance. Simultaneously, it strives to decrease the similarity between all negative pairs, which include all other utterances within the mini-batch or memory bank. Through the contrastive learning process, a model can effectively capture the speaker information without requiring additional adaptation. However, the potential presence of false negative samples (i.e., utterance pairs from the same speaker) may cause instability during the training process, ultimately leading to performance degradation.

To address this limitation, recent research has explored non-negative frameworks that remove the need for negative samples [12, 13]. One of the representative non-negative frameworks is distillation with no label (DINO) [13], a kind of self-distillation method which utilizes a teacher-student architecture [14]. DINO involves updating the teacher network with an exponential moving average (EMA) of the student network's parameters. The main objective of DINO framework is to minimize the discrepancy between the output features generated by the teacher and student networks given different cropped and augmented views of the same input. In application to SSL-based speaker embedding learning, DINO-based frameworks [15–19] have demonstrated superior discriminability compared to traditional contrastive learning-based frameworks, such as SimCLR and MoCo [8, 16]. Nevertheless, the DINO frameworks focus primary on the global view (i.e., the utterance-level), which may not sufficiently consider the inherent structure within the data [20].

The inherent structure of speech encompasses the unique characteristics of the speaker, ranging from physical properties (e.g., fundamental frequency and harmonics) to phonetic properties (e.g., timbre and intonation), which are vital for enriching the speaker representation [21–24]. To leverage these crucial factors, this study aims to improve DINO-based SSL framework for SV by incorporating a mask modeling approach. We propose the relational mask prediction (RMP) loss to leverage the inherent inter-frame structure of speech. The proposed loss encourages the model to learn the relationships between speech features rather than focusing solely on predicting the masked values themselves. Additionally, we introduce the block aggregation Transformer (BA-Transformer) architecture which is advantageous for enriching frame-level features, compared to recent SSL-based SV systems that exploit time-delay neural networks (TDNN) or convolutional neural networks (CNN) [11, 16]. By utilizing RMP loss and BA-Transformer structure, the proposed framework effectively extracts speaker embeddings by understanding the structure of the speech data.

The models were trained on the VoxCeleb2 development dataset omitting speaker labels. We investigated the initial speaker representation without an iterative clustering method [16]. Experimental results showed that the proposed framework outperforms recent SSL-based SV models and achieves superior generalizability across various evaluation datasets. Furthermore, we demonstrate the effectiveness of the proposed approach through ablation experiments.

2. DINO-based Self-Supervised SV

DINO [13], originally proposed for the field of computer vision, has been successfully applied to the SSL-based SV scenario demonstrating competitive performance [15, 16]. DINO is a non-negative self-distillation framework that exploits the mean-teacher strategy. Consequently, a student model acquires knowledge from the corresponding teacher model, which consists of an EMA of the student's own parameters. The framework organizes global and local views from the input using multi-crop and various distortion augmentation techniques. The previous SV literature has designated global views as long utterances (U_{lo}) and local views as short

*Equal contribution

10.21437/Interspeech.2024-362

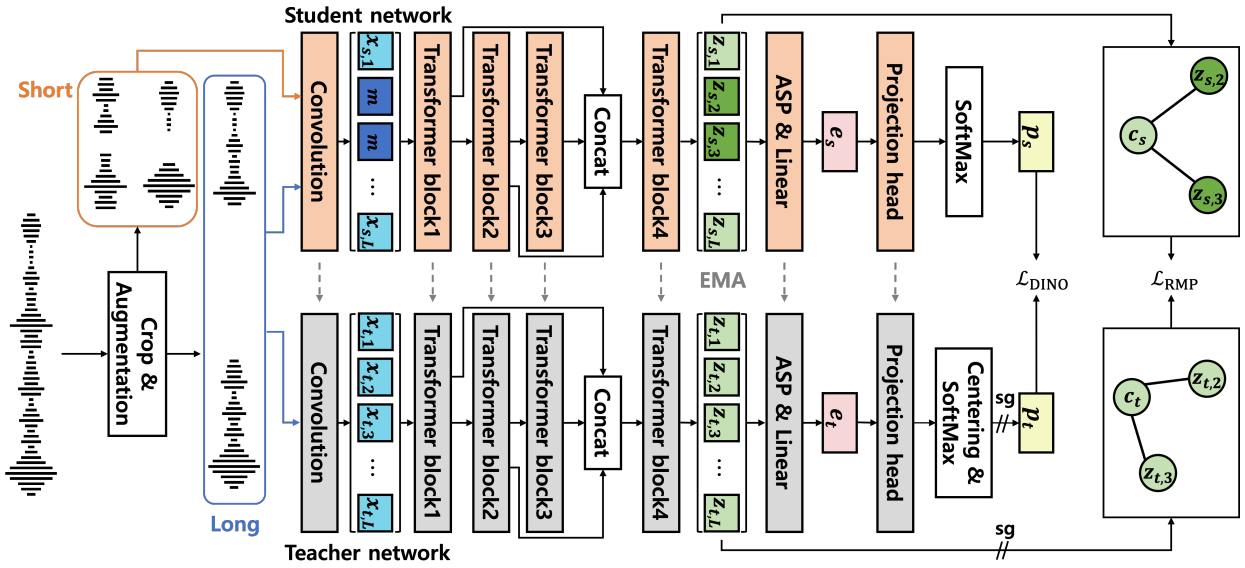


Figure 1: *An overview of the proposed framework.* **EMA** *refers to the exponential moving average, and* **sg** *represents the stop gradient operation.*

utterances (U_{sh}) [15, 16]. All utterance sets are fed into the student model, while only the long utterance set is fed into the teacher model. The student network is trained to match the output probability distribution of the teacher, by reducing the following loss:

$$\mathcal{L}_{\text{DINO}} = \sum_{u \in U_{lo}} \sum_{\substack{v \in U_{lo} \cup U_{sh} \\ v \neq u}} H(p_t(u), p_s(v)), \tag{1}$$

where $H(\cdot,\cdot)$ is the cross entropy function. $p_t(\cdot)$ and $p_s(\cdot)$ represent the output distributions of teacher and student networks, respectively. The non-negative framework requires specific preventive measures to avoid collapsing into trivial solutions [12]. To achieve this, DINO applies a centering and sharpening method to the output of the teacher and student networks, $g_t(\cdot)$ and $g_s(\cdot)$, as follows:

$$\begin{aligned} p_t(u) &= \text{softmax}((g_t(u) - c)/\tau_t), \\ p_s(v) &= \text{softmax}(g_s(v)/\tau_s), \end{aligned} \tag{2}$$

where c is the center of the teacher's distribution for centering, and τ_t and τ_s are the temperatures of the teacher and student networks for sharpening.

3. Proposed Approach

Masked modeling, also known as mask prediction, is an SSL method that obscures portions of data and then prompts the model to infer the concealed information using the unmasked data [25]. Through this training strategy, a model can represent internal relationships of data and capture long-term information [26, 27]. This study proposes the integration of mask modeling into DINO framework to enhance speaker discriminability by considering the inherent inter-frame structure of speech.

3.1. Block aggregation Transformer

For accurate mask prediction training, it is crucial that a model considers global context information at the frame level [25, 27]. Traditional convolutional operation-based architectures such as TDNNs and CNNs, which are commonly used in SV, struggle to model global frame relationships due to their inherent locality bias [2, 4, 28]. In contrast, Transformers [29] effectively capture global information across a sequence via self-attention mechanism. Therefore, we have incorporated Transformers into the proposed framework and, to the best of our knowledge, this is the first study to use a Transformer-based architecture for an SSL-based speaker embedding extraction.

Furthermore, we hypothesized that capturing powerful global information would lead to more effective mask modeling. Aggregating features from all blocks into a final pooling layer is a well-known scheme to derive discriminative utterance-level representations [28, 30]. Inspired by this method, we propose a BA-Transformer that combines features from all preceding blocks in the final block for enriching frame-level features. The following equation describes the aggregation process based on the student model:

$$Z_s = O_{s,4} = f_{s,4}(\text{Concat}(O_{s,1}, O_{s,2}, O_{s,3})). \tag{3}$$

$O_{s,i} \in \mathbb{R}^{B \times L \times D}$ represents the output from the i-th student's Transformer block $f_{s,i}$, where B, L, and D represent the batch size, frame length, and dimension, respectively. $Z_s = [\mathbf{z}_{s,1}, \mathbf{z}_{s,2}, ..., \mathbf{z}_{s,L}] \in \mathbb{R}^{B \times L \times (D \times 3)}$ denotes the output of the student's final block.

The final block's output, Z_s, is used directly for mask modeling and is converted into speaker embeddings, $\mathbf{e}_s$, via the attentive statistics pooling (ASP) [31] and a linear layer. This embedding extraction process is identical for the teacher model, and we used the embeddings extracted from the teacher model in the evaluation.

3.2. Relational mask prediction

The inherent structure of speech contains unique speaker characteristics, such as pitch and intonation, which are important for discriminative speaker representation [21–24]. To explicitly leverage this structural information of speech for SSL-based embedding learning, we propose the RMP loss. The RMP loss encourages the student network to predict the cosine similarity between masked and unmasked frame features using the guidance provided by the teacher network in the self-distillation framework. This loss function simplifies the training process by eliminating the need for separate pseudo-label generation techniques such as vector quantization or MFCC clustering [27, 32], and provides task-appropriate self-supervision in online.

Let us denote the convolution layer's output of the student network for input u as $X_s^u = [\mathbf{x}_{s,1}^u, \mathbf{x}_{s,2}^u, ..., \mathbf{x}_{s,L}^u]$. We generate a set of k random mask indices $M \subset [L]$, and replace the features corresponding to the mask indices with a single learnable mask embedding ($\mathbf{m}$) as in [27]. The corrupted version of X_s^u is represented as $\tilde{X}_s^u$ and the

Table 1: *Configuration of the input utterance sets for the proposed framework. DINO loss is computed over the entire combination of output distributions excluding the augmented input from the same long utterance, while RMP loss is computed only for the pairs of frame-level features for the same long utterance.*

Input set	Configuration		Model		Training target	
	Dur	Num	Teacher	Student	$\mathcal{L}_{\text{DINO}}$	$\mathcal{L}_{\text{RMP}}$
Long (U_{lo})	4s	2	✓	✓	Diff pair	Same pair
Short (U_{sh})	2s	4	×	✓	Long	×

teacher is given the unmasked features X_t^u. The masked $\tilde{X}_s^u$ and the intact X_t^u are processed into frame-level hidden features $\tilde{Z}_s^u$ and Z_t^u through Transformer blocks within the student and teacher networks, respectively. The relational matrix ($R \in \mathbb{R}^{B\times k}$) consists of the cosine similarities between the centroid of unmasked features and the masked features to represent the feature's relationship, as follows:

$$R_s^u = \text{sim}(C_s^u, \tilde{Z'}_s^u), \tag{4}$$

$$C_s^u = \frac{1}{L-k} \sum_{\substack{i=1 \\ i \notin M}}^{L} \tilde{z}_{s,i}^u, \tag{5}$$

where $\tilde{Z'}_s^u \in \mathbb{R}^{B\times k\times (D\times 3)}$ is the subset of $\tilde{Z}_s^u$ corresponding to the mask index, and $C_s^u \in \mathbb{R}^{B\times 1\times (D\times 3)}$ is the centroid of the features of the unmasked index. For cosine similarity computations ($\text{sim}(\cdot,\cdot)$), the centroid is broadcast along the frame axis. The teacher network's relational matrix (R_t^u) is also generated using the student network's mask index to build the correct relational self-supervision.

Finally, the RMP loss is calculated as the difference between the student and teacher relational matrices via the mean squared error $E(\cdot,\cdot)$, as follows:

$$\mathcal{L}_{\text{RMP}} = \sum_{u\in U_{lo}} E(R_t^u, R_s^u). \tag{6}$$

By minimizing the RMP loss, the student network learns to understand the inherent structure of the speech data through the prediction of relationships between masked and unmasked features. Compared to direct reproduction of a teacher's output, this indirect approach promotes feature diversity, potentially leading to improved speaker representability and discriminability.

3.3. DINO framework with RMP loss

Figure 1 depicts an overview of the proposed DINO-based SV framework with BA-Transformer and RMP loss, and Table 1 summarizes the input configuration and training targets of the framework. We construct two cropped utterance sets with various augmentations (U_{lo} and U_{sh}) for a single utterance. Then, the teacher model processes only the long utterance sets and the student model handles the full set of utterances. The student network is trained to minimize both DINO and RMP losses (equations (1) and (6)) using output distributions and the processed frame-level features, respectively, as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{DINO}} + \lambda \mathcal{L}_{\text{RMP}} \\ &= \sum_{u\in U_{lo}} \sum_{\substack{v\in U_{lo}\cup U_{sh} \\ v\neq u}} H(p_t(u), p_s(v)) + \lambda E(R_t^u, R_s^u),\end{aligned} \tag{7}$$

where λ is the weight for the RMP loss. DINO loss is optimized for different input utterance pairs to avoid training collapse, while RMP loss is optimized for the same input utterance pairs to learn the internal structure of the data. Thus, the combination of DINO and RMP loss is complementary and can extract more discriminative speaker embeddings.

4. Experimental setup

4.1. Dataset and evaluation metric

Our frameworks were trained using the VoxCeleb2 [34] development set, which contains 1,092,009 utterances, without labels. Primarily, the performance of the frameworks was evaluated using the VoxCeleb1 original test dataset (Vox1-O). In addition, we investigated the generalization ability of the model through VOiCES development [35], VoxSRC-22 development [36] and CN-Celeb evaluation [37] sets, including extended and hard trials on the VoxCeleb1 dataset (Vox1-E, -H) [34].

The embeddings derived from the teacher model were used to compute the cosine similarity score. We compared the frameworks based on the equal error rate (EER) and the minimum detection cost function (C_{det}^{min}). C_{det}^{min} was calculated by setting $P_{target} = 0.05$ and $C_{miss} = C_{fa} = 1$.

4.2. Implementation details

Input features. We employed an 80-dimensional log Mel-spectrogram as input features. These acoustic features were extracted using a 512-point FFT and a Hamming window with a width of 25ms and a hop size of 10ms. We also applied instance normalization [38] to the features.

Data augmentation. The long utterance set, U_{lo}, consists of two 4-second segments, and the short utterance set, U_{sh}, consists of four 2-second segments. All utterance sets were augmented using MUSAN [39] with SNR ranging from 0dB to 20dB and applying the room impulse response.

Training configuration of RMP loss. Specific features of the output from the convolutional layer, which corresponds to the student's long utterance input, were replaced with a 382-dimensional learnable mask embedding. For effective learning of the proposed framework, we gradually increased the difficulty of RMP loss by boosting the masking ratio from 5% to 10% during the training process. We set the loss weight of RMP (λ) to 1.

4.3. Hyper-parameters.

Models were trained using AdamW [40] optimizer with a mini-batch of 512 during 300 epochs. We linearly warmed up the learning rate (LR) to 1e-3 for the first 10 epochs, then gradually decreased it to 1e-8 over the subsequent 290 epochs using a cosine LR scheduler [41]. The BA-Transformer structure comprised three Transformer blocks with a hidden size of 384 and the last Transformer block with a hidden size of 1,152. The speaker embedding had 512 dimensions, and the projection head consisted of linear layer with weights of size 512 $\times$ 65,536. For sharpening, we set the teacher's temperature to 0.04 and the student's temperature to 0.1. The rest of the hyper-parameters follow the settings in [13].

Table 2: *Performance comparison between the recently proposed SSL-based SV systems, reported in terms of EER (%) and MinDCF (C_{det}^{min}) on the Vox1-O.*

Method	Architecture	EER (%)	C_{det}^{min}
MoCo + WavAug [8]	TDNN	8.23	0.5900
Contrastive [16]	ECAPA-TDNN	7.36	N/R
SSReg [11]	Thin-ResNet34	6.99	0.4340
DINO [18]	ECAPA-TDNN	6.70	0.4116
DINO-waveform [33]	RawNet3	5.40	0.3396
DINO + CL [18]	ECAPA-TDNN	4.47	0.3057
RDINO [19]	ECAPA-TDNN	3.29	0.2470
DINO + RMP (Ours)	**BA-Transformer**	**2.62**	**0.1736**

Table 3: *Performance evaluation (EER (%) / C_{det}^{min}) of SSL-based SV systems with different methods and architectures on various test sets, including in-domain and out-of-domain scenarios.*

Sys	Architecture	# params	Mask prediction	In-domain			Out-of-domain		
				Vox1-O	Vox1-E	Vox1-H	VOiCES	VoxSRC-22	CN-Celeb
#1	ECAPA-TDNN	26.3M	×	4.08 / 0.2301	4.61 / 0.2821	9.36 / 0.5170	2.55 / 0.2796	10.25 / 0.6059	15.91 / 0.6681
#2	Vanilla Transformer	24.9M	×	3.52 / 0.2456	3.81 / 0.2610	6.99 / 0.4051	2.15 / 0.1749	9.20 / 0.5199	15.03 / 0.6892
#3	BA-Transformer	25.7M	RMP	**2.62 / 0.1736**	**2.84 / 0.1936**	**5.14 / 0.3036**	**1.66 / 0.1374**	**7.51 / 0.4748**	**12.67 / 0.5775**

Table 4: *Experimental results of the ablation study for the proposed approach applied to the DINO-based framework on the Vox1-O trial.*

Sys	Architecture	Mask prediction	EER (%)	C_{det}^{min}
#4	Vanilla Transformer	×	3.52	0.2456
#5	BA-Transformer	×	3.02	0.2156
#6	BA-Transformer	MSE	2.88	0.1869
#7	BA-Transformer	RMP	**2.62**	**0.1736**

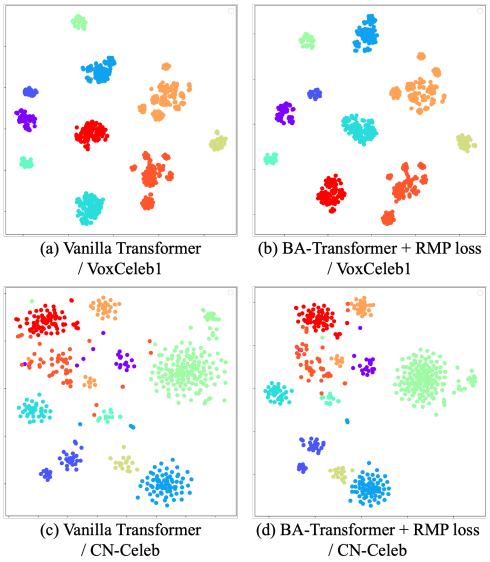
5. Results

Table 2 presents the comparative evaluation results of various SSL-based SV systems. Analyzing the results of recent studies, it is clear that non-negative DINO-based frameworks [18, 19, 33] outperform contrastive learning-based methods [8, 11, 16]. The proposed framework achieved a remarkably competitive EER of 2.62%, demonstrating a relative improvement of 20.36% over RDINO [19], the most advanced DINO-based system.

Table 3 shows the comprehensive evaluations, including out-of-domain datasets, to measure the generalization performance without any additional training for adaptation. System #1 represents a DINO framework using the ECAPA-TDNN structure [28], with the number of parameters scaled to match Transformer's model capacity for a fair comparison. As a baseline, we constructed a vanilla Transformer (system #2) with 9 Transformer blocks without any aggregation scheme. When comparing systems #1 and #2, Transformer model exhibited a performance enhancement with an average relative error reduction (RER) of 18.8% in the in-domain evaluation. However, the performance gains in out-of-domain evaluations were relatively limited, exhibiting only a 5.53% improvement on the CN-Celeb evaluation set. In contrast, our proposed framework demonstrated significant performance improvements in both the in-domain and out-of-domain evaluations, yielding RERs of 39.75% and 27.23%, respectively. These results provide substantial evidence for the superior generalization capabilities of our proposed framework.

We conducted ablation experiments to analyze the effectiveness of the proposed approaches, and the results are outlined in Table 4. Systems #4 and #7 are identical to systems #2 and #3 in Table 3, respectively. The system #5 was trained using only DINO loss without mask prediction to investigate the effectiveness of the BA-Transformer structure. We also designed a framework using MSE loss-based mask prediction (system #6), which directly predicts the teacher's output corresponding to each frame index, for comparison with the RMP loss. System #5 outperforms system #4, demonstrating the effectiveness of the BA-Transformer which further exploits frame-level features. The comparative analysis of the performance of systems #5, #6, and #7 showed that incorporating mask prediction learning can produce more discriminative speaker embeddings by learning the internal relations of the speech. We found that learning the relational information between frames through the proposed RMP loss is a more effective mask modeling method than simply reconstructing the teacher's output (#6 vs. #7). These results highlight the significance of the structure of BA-Transformer and RMP loss.

In Figure 2, we visualized the self-supervised speaker embeddings to explore the representability in different evaluation domains.



(a) Vanilla Transformer / VoxCeleb1 (b) BA-Transformer + RMP loss / VoxCeleb1

(c) Vanilla Transformer / CN-Celeb (d) BA-Transformer + RMP loss / CN-Celeb

Figure 2: *Visualizations of speaker embeddings using t-SNE. The embeddings are derived from the vanilla Transformer and the proposed framework (BA-Transformer + RMP loss) on the VoxCeleb1 test set and the CN-Celeb evaluation set.*

Specifically, we arbitrarily selected 10 speakers from the VoxCeleb1 test set (Figure 2-(a) and (b)) and the CN-Celeb evaluation set (Figure 2-(c) and (d)). Figure 2-(a) and (c) display the results from the vanilla Transformer (system #2 in Table 3), while Figure 2-(b) and (d) exhibit the results from the proposed framework (BA-Transformer + RMP loss). Comparing Figure 2-(a) and (b), it is evident that both the vanilla Transformer and the proposed framework yield comparable levels of speaker representation on the VoxCeleb1 test dataset. On the contrary, in the CN-Celeb evaluation dataset, the embeddings generated from our proposed framework demonstrate greater discriminability (Figure 2-(c) vs. (d)). This is particularly noticeable in terms of lower intra-class variance compared to the embeddings of vanilla Transformer. We infer that the proposed framework improves the compactness of speaker embeddings by capturing internal relational information from the speech even in out-of-domain scenarios.

6. Conclusion

In this study, we present an enhanced SSL-based SV framework, incorporating a mask modeling. We propose a relational mask prediction loss to capture the internal structural information of speech and introduce a block aggregation Transformer to enrich frame-level features. Our proposed framework outperforms recently reported models and demonstrates strong generalizability in both in-domain and out-of-domain scenarios. Ablation studies confirm the effectiveness of both the BA-Transformer and the RMP loss. We trained the model without labels, but the training dataset includes the assumption that each utterance contains only one speaker. As future work, we aim to scale the framework by leveraging unlabeled in-the-wild data.

7. References

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4052–4056.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[3] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[4] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[5] D. Alexey, P. Fischer, J. Tobias, M. R. Springenberg, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE TPAMI*, vol. 38, no. 9, pp. 1734–1747, 2016.

[6] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[7] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, 2016, pp. 69–84.

[8] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6723–6727.

[9] H. Zhang, Y. Zou, and H. Wang, "Contrastive self-supervised learning for text-independent speaker verification," in *International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6713–6717.

[10] J. Kang, J. Huh, H. S. Heo, and J. S. Chung, "Augmentation adversarial training for self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1253–1262, 2022.

[11] M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan, "Self-supervised speaker verification with simple siamese network and self-supervised regularization," in *International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6127–6131.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS 2014*, 2015.

[15] J. Cho, J. Villalba, and N. Dehak, "The jhu submission to voxsrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.

[16] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, "Self-supervised speaker recognition with loss-gated learning," in *International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6142–6146.

[17] C. Zhang and D. Yu, "C3-dino: Joint contrastive and non-contrastive self-supervised learning for speaker verification," *Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1273–1283, 2022.

[18] H.-S. Heo, J. weon Jung, J. Kang, Y. ki Kwon, B.-J. Lee, Y. J. Kim, and J. S. Chung, "Curriculum Learning for Self-supervised Speaker Verification," in *INTERSPEECH*, 2023, pp. 4693–4697.

[19] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, "Pushing the limits of self-supervised speaker verification using regularized distillation framework," in *International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[20] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *International Conference on Learning Representations*, 2022.

[21] B. Zellner, "Pauses and the temporal structure of speech," in *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.* John Wiley, 1994, pp. 41–62.

[22] C. Gussenhoven, "The phonology of tone and intonation," 2004.

[23] W.-G. Gong, L.-P. Yang, and D. Chen, "Pitch synchronous based feature extraction for noise-robust speaker verification," in *2008 Congress on Image and Signal Processing*, vol. 5, 2008, pp. 295–298.

[24] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech communication*, vol. 50, no. 10, pp. 782–796, 2008.

[25] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[26] S. Chen, Y. Wu, C. Wang, S. Liu, Z. Chen, P. Wang, G. Liu, J. Li, J. Wu, X. Yu *et al.*, "Why does self-supervised learning for speech recognition benefit speaker recognition?" *Interspeech*, 2022.

[27] 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.

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

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] J.-h. Kim, H.-j. Shim, J. Heo, and H.-J. Yu, "Rawnext: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies," in *International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7647–7651.

[31] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Interspeech*, 2018.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[33] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," *Interspeech*, 2022.

[34] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[35] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices Obscured in Complex Environmental Settings (VOiCES) Corpus," in *INTERSPEECH*, 2018, pp. 1566–1570.

[36] J. Huh, A. Brown, J.-w. Jung, J. S. Chung, A. Nagrani, D. Garcia-Romero, and A. Zisserman, "Voxsrc 2022: The fourth voxceleb speaker recognition challenge," *arXiv preprint arXiv:2302.10248*, 2023.

[37] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *nternational Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7604–7608.

[38] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

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

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[41] ——, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2016.