Interspeech 2024
1-5 September 2024, Kos, Greece

Improving Noise Robustness in Self-supervised Pre-trained Model for Speaker Verification

Chan-yeong Lim, Hyun-seo Shin*, Ju-ho Kim, Jungwoo Heo, Kyo-Won Koo, Seung-bin Kim, Ha-Jin Yu†*

University of Seoul, Republic of Korea

cksdud585@naver.com, gustjtls123@naver.com, wngh1187@naver.com, jungwoo4021@gmail.com, kkwr0504@naver.com, kimho1wq@naver.com, hjyu@uos.ac.kr

Abstract

Adopting self-supervised pre-trained models (PMs) in speaker verification (SV) has shown remarkable performance, but their noise robustness is largely unexplored. In the field of automatic speech recognition, additional training strategies enhance the robustness of the models before fine-tuning to improve performance in noisy environments. However, directly applying these strategies to SV risks distorting speaker information. We propose a noise adaptive warm-up training for speaker verification (NAW-SV). The NAW-SV guides the PM to extract consistent representations in noisy conditions using teacher-student learning. In this approach, to prevent the speaker information distortion problem, we introduce a novel loss function called extended angular prototypical network loss, which assists in considering speaker information and exploring robust speaker embedding space. We validated our proposed framework on the noise-synthesized VoxCeleb1 test set, demonstrating promising robustness.

Index Terms: speaker verification, self-supervised learning, teacher-student learning, noisy environments

1. Introduction

Self-supervised learning (SSL) techniques, which utilize large amounts of unlabeled data to enable models to derive effective speech representations, have become mainstream in the speech signal processing field. Building on the success of SSL, various researchers in the speech domains have adopted self-supervised pre-trained models (PMs) such as Wav2Vec 2.0 [1], HuBERT [2], WavLM [3] in their respective domains, leading to the achievement of remarkable performance. Speaker verification (SV), which is the task of verifying that the speaker of an input speech is the speaker enrolled in the system, is one of the main areas where PM has been successfully utilized [3, 4].

While PMs have demonstrated notable success in SV tasks, potential challenges could compromise their performance. Numerous studies have observed that PM-based systems outperform in trained in-domain environments but struggle significantly in noisy out-of-domain conditions [5–8]. Given that most PM-based SV studies have applied PMs without addressing this degradation issue, these systems may encounter similar performance declines in noisy environments. Investigating these issues is crucial for the practical application of PM-based SV systems in real-world scenarios, typically in noisy environments.

In automatic speech recognition (ASR), several studies have attempted to solve this problem by using additional training before fine-tuning the target task. These studies designed various training strategies to make the PMs robust to noise during the warm-up training phase, as shown in (b) in Figure 1. To this end, Wang *et al.* [5] and Zhu *et al.* [9] employed contrastive learning strategies to ensure consistent performance across clean and noisy speech conditions. Another approach, HuBERT-AGG [10], demonstrated the viability of constructing a noise-robust PM by distilling aggregated noise-invariant features related to ASR. These studies have proposed various approaches with the aim of enabling the model to extract noise-invariant representations.

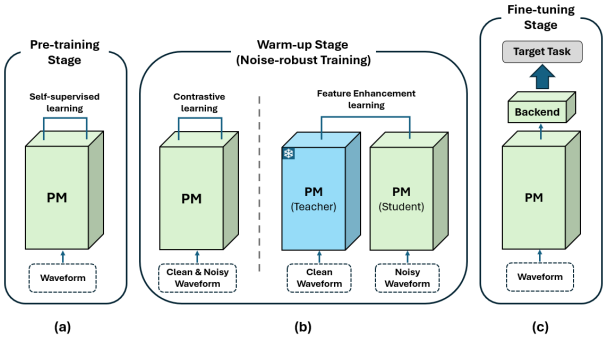

Figure 1: *A simplified illustration of a three-step learning approach for constructing a noise-robust pre-trained model (PM).*

However, some considerations should be made when applying these methods to SV. According to the study by Kolboek *et al.* [11], feature enhancement (FE) learning, which excludes noise information in the speech using a loss function such as mean squared error loss (MSE), could potentially distort speaker information. This has been observed in several other SV studies, which mitigated this problem by conducting FE learning that takes into account speaker information [12–14]. Therefore, to build a noise-robust PM that reduces the distortion of speaker information, we assert that speaker information should be factored in from the warm-up training phase.

With this objective, we introduce a warm-up training method tailored explicitly for SV, called noise adaptive warm-up training for speaker verification (NAW-SV). The NAW-SV method trains a PM for noise robustness through three key components: 1) teacher-student (TS) learning, 2) extended angular prototypical network loss (E-APN), and 3) Adapter [15]. TS learning, adopted for its effectiveness in ASR, encourages a student model to extract representations similar to the teacher's, thereby serving as a potentially promising method for constructing noise-resistant PMs for SV. To prevent the issue of speaker information distortion, we introduced E-APN in the warm-up training phase, a novel metric loss designed to explore the noise-robust speaker embedding space. Also, it may be more chal-

*Equal contribution
†Corresponding author.

10.21437/Interspeech.2024-1630

lenging for the model to output an expression identical to the clean speech from the noisy speech while considering speaker information. Therefore, to increase the model's capacity, we have also added an adapter, which can effectively change the distribution of the transformer structure with fewer parameters. Through these proposed techniques, we believe the PM will be robustly constructed for SV by performing noise compensation learning while appropriately considering speaker information.

For evaluation, we used the VoxCeleb1 [16] & 2 [17] dataset and the MUSAN [18] dataset. The experiments show that our proposed framework achieves significant performance improvements over the existing PMs in noisy environments.

2. Proposed methods

2.1. Architecture

The warm-up training strategy has been substantiated as capable of enhancing resilience and maintaining performance stability within noisy environments [5, 9, 10]. In line with this approach, we propose a novel approach called noise adaptive warm-up training for speaker verification (NAW-SV) using TS learning with adapters. As shown in Figure 2, the NAW-SV method leverages a PM as the teacher model, keeping its parameters fixed. The student model maintains the same structure as the teacher model, and its parameters are initially set identically to the teacher's. This model incorporates a unique element, the 'adapter', in its encoder layers. The adapter handles the increased complexity of the noise-robust task assigned to the student, utilizing this model during the fine-tuning phase. Through this structure, the student model can potentially possess robustness against noise for the target task.

For TS learning and the use of metric loss, the mini-batch is configured to present clean utterances to the teacher and both clean and noisy utterances to the student. This configuration, as shown in Equation 1, makes the system preserve the performance of clean utterances within the system by mapping the features extracted from clean inputs to the teacher's output [14].

$$M_t = [U_1^1, U_2^1, ..., U_n^1, \quad M_s = [U_1^1, U_2^1, ..., U_n^1, \\ U_1^2, U_2^2, ..., U_n^2], \qquad \tilde{U}_1^2, \tilde{U}_2^2, ..., \tilde{U}_n^2], \tag{1}$$

Here, M_t and M_s are the input mini-batches for the teacher and student, respectively. U_i^j represents the j-th clean utterance from the i-th speaker, while $\tilde{U}$ signifies a noisy utterance, and n corresponds to the number of speakers for a mini-batch.

In the teacher model, M_t is processed to the feature maps ($h_t^{(l)}$, $l \in \{0, 1, ..., L\}$), encompassing the convolutional layer (f_t) and L transformer layers ($g_t^{(l)}$) as Equation 2.

$$h_t^{(0)} = f_t(M_t), \quad h_t^{(l+1)} = g_t^{(l+1)}(h_t^{(l)}), \tag{2}$$

In contrast, the student model takes M_s as an input, processes it through the convolutional layer (f_s), and applies masking to both the temporal and feature axes. We posit that this masking technique could potentially assist the model in identifying the inherent structure and context of the audio sequence. By predicting the masked segments and compensating for utterances affected by noise, the model's resilience against noise interference could be significantly enhanced. After that, the outputs, $h_s^{(0)}$, are passed through the student's encoder layers with the adapters, which are composed of a downsampling layer ($W_{down} \in \mathbb{R}^{768\times64}$), an ELU [19] activation function ($\sigma(\cdot)$), and an upsampling layer ($W_{up} \in \mathbb{R}^{64\times768}$).

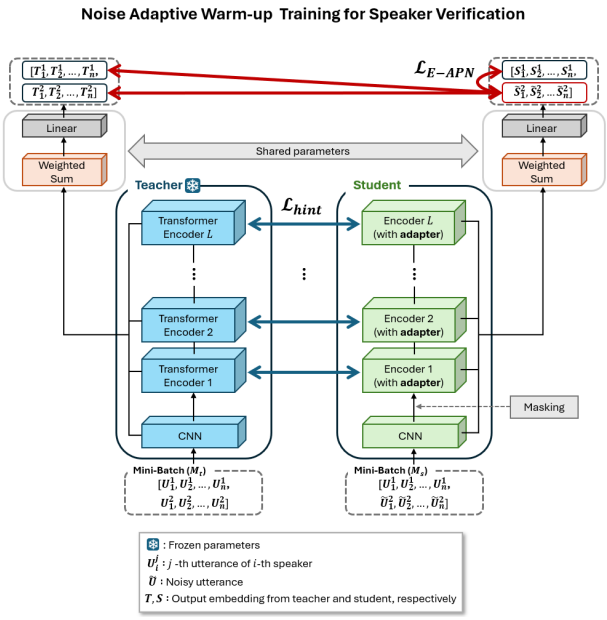

Figure 2: *Illustration of the proposed NAW-SV method.*

Processing the $h_s^{(0)}$ to $h_s^{(l)}$ can be represented by the following equation:

$$h_s^{(0)} = \text{Mask}(f_s(M_s)), \tag{3}$$

$$z_s^{(l+1)} = \text{LayerNorm}(h_s^{(l)} + \text{MHSA}(h_s^{(l)})), \tag{4}$$

$$\text{Adpt}(z_s^{(l+1)}) = \sigma(z_s^{(l+1)} W_{down}) W_{up}, \tag{5}$$

$$h_s^{(l+1)} = g_s^{(l+1)}(h_s^{(l)}) \\ = \text{LayerNorm}(z_s^{(l+1)} + \text{FFN}(z_s^{(l+1)}) + \text{Adpt}(z_s^{(l+1)})), \tag{6}$$

where $\text{Mask}(\cdot)$ symbolizes the masking process, $\text{MHSA}(\cdot)$, $\text{FFN}(\cdot)$, and $\text{Adpt}(\cdot)$ denote multi-head self-attention layer, feed-forward network layer, and the adapter, respectively. The term $z_s^{(l)}$ represents the output that has passed through $\text{LayerNorm}(\cdot)$ and $\text{MHSA}(\cdot)$ of the l-th layer with the residual path.

2.2. Loss functions

The NAW-SV aims to enable the PM to extract noise-invariant representations for SV. To this end, we implemented MSE to match the student and the teacher model outputs closely. The student model receives clean-noisy pair utterances as input, whereas the teacher model only uses clean utterances. The loss reduces the Euclidean distance between these different outputs $h_s^{(l)}$ and $h_t^{(l)}$ as follows:

$$\mathcal{L}_{hint} = \text{MSE}(h_t^{(L)}, h_s^{(L)}) + \alpha \sum_{l=1}^{L-1} \text{MSE}(h_t^{(l)}, h_s^{(l)}), \tag{7}$$

where α is a value set to less than 1 to ensure that the knowledge transfer from the teacher to the student remains within appropriate limits. By minimizing $\mathcal{L}_{hint}$, the student is trained to extract a consistent representation over noise.

However, in the SV task, compensating for noise only may pose a risk of distorting speaker information [12–14]. To mitigate this distortion, we propose a novel metric loss, extended angular prototypical network loss (E-APN), which is a modification of the angular prototypical network loss (APN) [22].

Table 1: *Comparison of EERs (%) of SV systems for noisy environments and self-supervised pre-trained models (PMs) in noise scenarios with various SNRs. ECAPA-TDNN (small) [20] was employed as the backend model for PMs.*

Training set	Model	EER (%)															
		Clean	Noise					Music					Babble				Avg
			0	5	10	15	20	0	5	10	15	20	0	5	10	15	20	
VoxCeleb1	VoiceID [12]	6.79	16.56	12.26	9.86	8.69	7.83	16.24	11.44	9.13	8.10	7.48	37.96	27.12	16.66	11.25	8.99	13.52
	Wu *et al.* [13]	7.60	13.12	10.57	9.28	8.59	8.10	12.92	10.10	8.95	8.35	7.95	20.11	12.02	9.63	8.48	7.99	10.24
	Cai *et al.* [21]	3.12	7.34	5.65	4.35	3.85	3.44	7.79	5.23	4.11	3.63	3.30	11.78	5.97	4.44	3.73	3.36	5.07
	ExU-Net [14]	2.76	6.80	5.23	4.07	3.39	3.10	7.35	4.90	3.69	3.14	2.93	9.57	5.52	4.06	3.28	2.99	4.55
	HuBERT Base	2.20	8.52	5.62	4.22	3.36	3.02	10.11	5.66	3.98	3.21	2.76	14.75	6.40	4.10	3.19	2.83	5.25
	HuBERT + NAW-SV	**1.89**	**6.87**	**4.28**	**3.27**	**2.79**	**2.54**	**7.15**	**4.13**	**3.15**	**2.55**	**2.29**	**11.34**	**4.88**	**3.28**	**2.67**	**2.42**	**4.09**
	WavLM Base+	1.56	6.08	3.82	2.84	2.40	2.22	6.76	3.90	2.64	2.23	1.98	9.05	4.03	2.83	2.26	1.97	3.54
	WavLM + NAW-SV	**1.45**	**5.30**	**3.35**	**2.44**	**2.01**	**1.81**	**5.52**	**3.39**	**2.35**	**2.00**	**1.69**	**7.14**	**3.59**	**2.53**	**2.05**	**1.75**	**2.96**
VoxCeleb2	HuBERT Base	1.24	6.23	3.70	2.60	2.07	1.69	7.05	3.57	2.49	1.77	1.62	11.16	4.41	2.55	1.85	1.61	3.48
	HuBERT + NAW-SV	**1.12**	**5.53**	**3.12**	**2.19**	**1.77**	**1.47**	**5.79**	**2.99**	**2.18**	**1.63**	**1.44**	**8.48**	**3.44**	**2.13**	**1.61**	**1.40**	**2.89**
	WavLM Base+	0.94	4.98	2.84	1.94	1.46	1.26	5,18	2.64	1.85	1.33	1.22	7.87	2.98	1.78	1.32	1.19	2.57
	WavLM + NAW-SV	**0.85**	**4.39**	**2.46**	**1.84**	**1.35**	**1.25**	**4.52**	**2.54**	**1.70**	**1.32**	**1.16**	**6.23**	**2.83**	1.84	1.45	**1.18**	**2.31**

We designed it to facilitate the exploration of a speaker embedding space that is robust against noise, which is in line with the TS learning strategy. The conventional APN reduces the intra-speaker variance and increases the inter-speaker variance by comparing the cosine similarity between embeddings within a mini-batch of the model. Thus, it assists in finding the optimal speaker embedding space. However, since the student model processes masked data, applying this loss directly has the potential to deviate from its intended functionality due to the missing information. To realize the objectives effectively, employing unmasked, clean embeddings from the teacher model would be valuable. Therefore, the E-APN utilizes not only the student's output but also the teacher's output, as follows:

$$C^1_{i,j} = w \cdot \cos(\tilde{S}^2_i, S^1_j) + b, \tag{8}$$

$$C^2_{i,j} = w \cdot \cos(\tilde{S}^2_i, T^1_j) + b, \tag{9}$$

$$C^3_{i,j} = w \cdot \cos(\tilde{S}^2_i, T^2_j) + b, \tag{10}$$

$$\mathcal{L}_{E-APN} = -\frac{1}{3n}\sum_{i=1}^{n}(\log\frac{e^{C^1_{i,i}}}{\sum_{j=1}^{n} e^{C^1_{i,j}}} + \log\frac{e^{C^2_{i,i}}}{\sum_{j=1}^{n} e^{C^2_{i,j}}} + \log\frac{e^{C^3_{i,i}}}{\sum_{j=1}^{n} e^{C^3_{i,j}}}), \tag{11}$$

where S and $\tilde{S}$ represent embeddings generated from two distinct input utterances - the clean and noisy utterances in the student's mini-batch. T is the embeddings generated by the teacher. These embeddings undergo a process via a fully connected layer following a weighted summation of hidden states from all layers ($h^{(l)}$, $l \in \{0, 1, ..., L\}$), with parameters shared between the teacher and the student, as shown in Figure 2. Continuing with the cosine similarity calculation, the metric $C^1_{i,j}$ is computed between $\tilde{S}^2_i$ and S^1_j, incorporating a learnable scale (w) and bias (b).

As optimization progresses, the E-APN mechanism aids in matching the noisy embeddings with the multiple clean embeddings from the teacher and the student, thereby fostering the development of more noise-robust space for distinguishing the speakers. Finally, our proposed NAW-SV method trains the PM by concurrently optimizing the two loss functions as follows:

$$\mathcal{L}_{NAW-SV} = \mathcal{L}_{hint} + \lambda\mathcal{L}_{E-APN} \tag{12}$$

3. Experiment setting

3.1. Dataset

To conduct a comparative analysis under noisy conditions, we employed the MUSAN [18] corpus as a noise source. We have partitioned the MUSAN corpus to prevent any overlap between training and test subsets and to ensure the incorporation of diverse noise types within each subset. During the NAW-SV phase, the training set of VoxCeleb1 [16] or VoxCeleb2 [17] dataset was utilized for the clean utterances, depending on the experiment. Noisy data was synthesized by adding the MUSAN data to the clean data with random SNR values ranging from 0 to 20. In the fine-tuning stage, the same VoxCeleb training set was used as input, and for data augmentation, the MUSAN training subset and RIR reverberation datasets [23] were employed. Throughout the evaluation phase, clean input was derived from the original VoxCeleb1 test utterances, while noisy input was simulated from datasets comprising SNRs 0, 5, 10, 15, 20 extracted from both the VoxCeleb1 test set and the MUSAN test subset. For out-of-domain noise analysis, Nonspeech100 dataset [24] was employed, with the evaluation conducted using configurations aligned with in-domain noise scenarios. The key metric is the equal error rate (EER), determined based on cosine similarity scores.

3.2. Pre-trained models and backend model

We verified the robustness of the proposed method in WavLM Base+ and HuBERT Base. These PMs were implemented through the HuggingFace transformer library [25]. The backend model we utilized was the ECAPA-TDNN (small), which processes the PM output through a weighted summation approach as proposed by Chen *et al.* [4].

3.3. Implementation details

In our experiment, the input was composed of utterances randomly cropped to 3 seconds. The Adam optimizer [26], without weight decay, was utilized across all experiments, and the initial learning rate was set at $10^{-5} \times 5$. During the NAW-SV phase, mini-batches comprise 192 clean and noisy utterances from 96 randomly chosen speakers. For the first 10 epochs, we only trained the adapters with the fixed parameters of the PM and then jointly trained for 40 epochs. The masking probability was set to 0.1 for the time axis and 0.05 for the feature axis, and we set α and λ to 0.1. During the fine-tuning phase, we used mini-batches composed of 24 samples. The AAM-softmax [27] was utilized as the speaker identification loss function with a margin of 0.15 and a scale of 20. In all experiments, this study did not use any voice activity detection technology or score normalization technology. Readers can access our code here [1].

4. Results

Table 1 provides a comparative analysis of recent SV systems and our proposed framework, which are designed for noisy en-

[1] https://github.com/chan-yeong0519/NAW-SV

Table 2: *Average EERs (%) for various types of noise under different training strategies, trained using the VoxCeleb1 dataset*

#Exp	Training strategy	Vox1 (EER %)				
		Clean	Noise	Music	Babble	Avg
#1	WavLM (Baseline)	1.56	3.47	3.50	4.03	3.54
#2	WavLM (increasing epoch)	1.67	3.57	3.55	3.85	3.53
#3	Joint training	2.80	5.25	5.22	5.54	5.18
#4	NAW-SV (w/o adapters)	1.57	3.12	3.12	3.59	3.17
#5	**NAW-SV**	**1.45**	**2.98**	**2.99**	**3.41**	**2.96**

vironments. In the VoxCeleb1 training setup, HuBERT outperforms other SV systems under clean conditions with an EER of 2.20%. Nonetheless, its effectiveness significantly diminishes in noisy situations. Conversely, WavLM, which incorporates denoising tasks during pre-training, surpasses existing SV systems in all evaluation conditions. These results show the potential of the PMs for SV but underscore the need to strengthen the robustness further. HuBERT and WavLM, trained via our proposed NAW-SV method, achieve relative error reduction rates (RERs) of approximately 22.1% and 16.4%, respectively, in average EER, demonstrating enhanced performance across all evaluation conditions. Furthermore, in the VoxCeleb2 training setup, HuBERT + NAW-SV achieves an average EER of 2.89%, marking an RER of approximately 17% compared to the original HuBERT. WavLM + NAW-SV records the best performance with an average EER of 2.31%, generally indicating a performance improvement over the original WavLM, even though it has been tasked with denoising. Through these results, we confirm our proposed approach's superiority to fortify the PMs against noise for SV.

Table 2 displays the results of the other learning strategies using the VoxCeleb1 dataset. In Experiment #2, we used 77 epochs instead of 27, and in experiment #3, the joint learning strategy is applied by integrating the NAW-SV and the fine-tuning stage. These experiments show their inability to tackle performance degradation, indicating that merely increasing the learning iteration or implementing learning strategies concurrently does not significantly enhance SV performance. To ascertain the impact of the adapters in our proposed framework, we conducted Experiment #4 in which the student was trained without the adapters. It improved performance in noisy conditions compared to the baseline using only the training strategy. However, the performance across all evaluation conditions fell short when contrasted with the NAW-SV. This finding emphasizes the effectiveness of adapters in enabling the model to handle noise-robust task, which is more challenging.

Table 3 is the experimental results of the loss function applied during the warm-up training stage. Experiment #7, which only employed MSE in the warming-up phase, recorded a performance decline in clean conditions; however, it demonstrated comparable performance to the baseline in noisy environments. These results suggest the potential of MSE-based TS learning in fostering robustness to noise while raising the possibility of distortion of speaker information. On the contrary, experiments applying APN (#8) and E-APN (#9) solely indicate a significant performance drop. We interpret these outcomes driven by the PM specializing in SV without preserving its original knowledge, thereby impairing its capacity to extract overall speech information. Therefore, it is essential to concurrently transmit original knowledge and improve noise restoration capability (MSE), as well as speaker-aware training (APN, E-APN).

Building upon these observations, Experiment #11, which employed both MSE and APN during the warm-up stage, offers an insightful perspective. This approach resulted in improved performance in noisy environments compared to the baseline.

Table 3: *Ablation experiments of the loss function for each type of noise. (Exp #6 is identical to Exp #1.)*

#Exp	Loss functions	Vox1 (EER %)				
		Clean	Noise	Music	Babble	Avg
#6	$\times$ (Baseline)	1.56	3.47	3.50	4.03	3.54
#7	MSE	1.93	3.82	3.57	4.18	3.77
#8	APN	3.09	5.42	5.27	5.50	5.25
#9	E-APN	2.41	4.46	4.41	4.79	4.42
#10	MSE & CCE	4.37	6.02	5.93	6.52	6.05
#11	MSE & APN	1.73	3.20	3.25	3.69	3.28
#12	MSE & E-APN	**1.45**	**2.98**	**2.99**	**3.41**	**2.96**

Table 4: *Experimental results from the synthesized VoxCeleb1 test set using an out-of-domain noise source (Nonspeech100).*

Noise type	SNR	HuBERT	HuBERT (**NAW-SV**)	WavLM	WavLM (**NAW-SV**)
Nonspeech	0	9.96	**8.07**	7.39	**6.19**
	5	6.07	**4.74**	4.18	**3.65**
	10	4.45	**3.46**	3.13	**2.65**
	15	3.47	**2.92**	2.52	**2.08**
	20	3.00	**2.54**	2.17	**1.88**
Average (EER %)		5.39	**4.35**	3.88	**3.29**

However, it is essential to note that replacing APN with categorical cross-entropy loss (CCE) in Experiment #10 led to the highest EER among all conducted experiments. This underlines the significance of the appropriate selection and combination of loss functions. Such a combination should allow the PM to learn both the consideration of speaker information and robustness to noise during the warm-up training phase without leaning towards one or the other. In contrast, the experiment that used CCE, which is focused on clearly distinguishing speakers, implies that the learning process concentrated more on identifying speaker information, leading to performance degradation. Experiment #12, which applies our proposed loss function, achieves the lowest EER and exhibits enhanced performance compared to Experiment #11. Through this, we verify the effectiveness of the E-APN, which operates to explore the noise-robust embedding space among various clean embeddings in our framework structure.

Table 4 displays the out-of-domain evaluation results of HuBERT- and WavLM-based SV systems. Both PMs, which perform additional noise-robust learning through the NAW-SV, achieve RERs of 19.3% and 15.2%, respectively, in average EER compared to their original. These experimental results confirm that our proposed method can enhance the generalizability and robustness of the out-of-domain noise data.

5. Conclusion

In this paper, we propose an additional training strategy, noise adaptive warm-up training for speaker verification (NAW-SV), to be applied prior to the fine-tuning stage for constructing noise-robust PMs for SV. The NAW-SV leverages teacher-student learning with adapters, enabling the extraction of noise-invariant consistent representations, thereby enhancing robustness against noise. Furthermore, we introduced a new metric loss, the extended angular prototypical network loss, which leads the PM to explore a noise-robust embedding space. Our proposed method demonstrated superior robustness to both in- and out-domain noise compared to the existing PM and also showed enhanced performance even in a clean environment. However, our proposed method still needs to be compared to other feature enhancement learning methods. We plan to address this in future research endeavors to enhance the robustness of the PMs further.

6. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (2023R1A2C1005744)

7. References

[1] 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.

[2] 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.

[3] 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*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[5] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7097–7101.

[6] Y. Masuyama, X. Chang, S. Cornell, S. Watanabe, and N. Ono, "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 260–265.

[7] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai, "A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[8] H. Sato, R. Masumura, T. Ochiai, M. Delcroix, T. Moriya, T. Ashihara, K. Shinayama, S. Mizuno, M. Ihori, T. Tanaka, and N. Hojo, "Downstream Task Agnostic Speech Enhancement with Self-Supervised Representation Loss," in *Proc. INTERSPEECH 2023*, 2023, pp. 854–858.

[9] Q.-S. Zhu, L. Zhou, J. Zhang, S.-J. Liu, Y.-C. Hu, and L.-R. Dai, "Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] W. Wang and Y. Qian, "Hubert-agg: Aggregated representation distillation of hidden-unit bert for robust speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] M. Kolboek, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 2016, pp. 305–311.

[12] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *Interspeech*, pp. 2888–2892, 2019.

[13] Y. Wu, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Joint feature enhancement and speaker recognition with multi-objective task-oriented network." *Interspeech*, pp. 1089–1093, 2021.

[14] J. Kim, J. Heo, H. Shim, and H. Yu, "Extended u-net for speaker verification in noisy environments," *Interspeech*, pp. 590–594, 2022.

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Interspeech*, pp. 2616–2620, 2017.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 1086–1090.

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

[19] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *4th International Conference on Learning Representations, ICLR*, Y. Bengio and Y. LeCun, Eds., 2016.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Interspeech*, pp. 3830–3834, 2020.

[21] D. Cai, W. Cai, and M. Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6469–6473, 2020.

[22] 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," *Interspeech*, pp. 2977–2981, 2020.

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[24] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations (ICLR)*, 2015.

[27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.