A STUDY OF JOINT FRAMEWORK FOR ROBUSTNESS AGAINST NOISE ON SPEAKER VERIFICATION SYSTEM

Sangwook Han, Youngdo Ahn, Kyeongmuk Kang, and Jong Won Shin

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology, Gwangju, Korea

ABSTRACT

As speaker verification (SV) technology is applied to various fields with the development of deep learning, robustness against noise is becoming more prominent. A common approach for suppressing the noise effect is to jointly train SV system along with a front-end enhancement module as pre-processing. However, in this paper, we explore a training method for a noise-robust system that focuses on enhancing the single SV system. In particular, we introduce the feature-robust loss to exploit the embeddings extracted from the pre-trained model, which is jointly trained with the enhancement module. Experimental results showed that our proposed method can mitigate the noise effect in various noisy conditions using only a single SV system, demonstrating its potential for further development.

Index Terms— speaker verification, speech enhancement, joint framework, noise robustness, security system

1. INTRODUCTION

Speaker verification (SV) is the task of authenticating the claimed identity of the enrolled speaker using an audio sample. In recent years, the performance of SV systems has significantly improved with the rapid development of deep neural networks (DNNs) to extract speaker-specific embedding [1–3]. At the same time, the demand for SV systems is increasing, and these systems are being widely used in many security applications such as smart door locks, IoT devices, and financial services.

While SV systems have made significant progress with clean audio recordings, the performance could be dramatically degraded in noisy and reverberation environments. In real-world scenarios, recorded speech signals are always interfered by various types of background noise and reverberations, which can increase both the false acceptance ratio (FAR) and false rejection ratio (FRR). Thus, ensuring noise robustness is crucial for SV systems especially when they are utilized for biometric authentication.

In order to mitigate the noise effect, there are two mainstream approaches: data augmentation and integrating an enhancement module. Firstly, SpecAugment [4] and multi-condition training (MCT) are the most prevalent data augmentation methods, which are commonly used to boost the robustness of SV systems. SpecAugment is a method for augmenting audio data, which utilizes time warping and masking techniques across both the frequency and time axes. For MCT, the SV system is trained with different kinds of signal-to-noise ratio (SNR) conditions, which helps improve its generalization ability and robustness. The second method is integrating the speech enhancement (SE) module at the front-end as pre-processing step. Generally, the SE module and back-end speaker classifier can be jointly trained to optimize the whole system. Since SE aims to improve the speech's perceptual quality and intelligibility by suppressing noise, it can be beneficial in reducing the impact of noise on SV systems. However, speech distortions caused by speech enhancement loss may lead to performance degradation in downstream tasks, particularly in scenarios with high SNR levels.

In this paper, we explore a joint training framework that aims to further enhance the single SV system to mitigate the impact of speech distortion in low SNR conditions. We utilize the combined model of SE and SV as a pre-trained model, employing it for the extraction of noise-robust embeddings. Moreover, we introduce the feature-robust loss, which can compensate for the shortcomings of the MCT-based single SV systems by utilizing the embeddings extracted from the pre-trained model. Experimental results confirmed that our proposed method can alleviate the noise effect in various noisy conditions compared to the MCT-based single SV system.

2. PROPOSED METHOD

We propose a joint training approach that focuses on alleviating the noise effect and speech distortion, enhancing a single SV system as shown in Figure 1. First, we pre-train the combined system $g_\xi(\cdot)$ with a set of weights ξ that consist of a speech enhancement network and a speaker verification net-

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2021-0-01835) supervised by the IITP (Institute of Information Communications Technology Planning Evaluation)

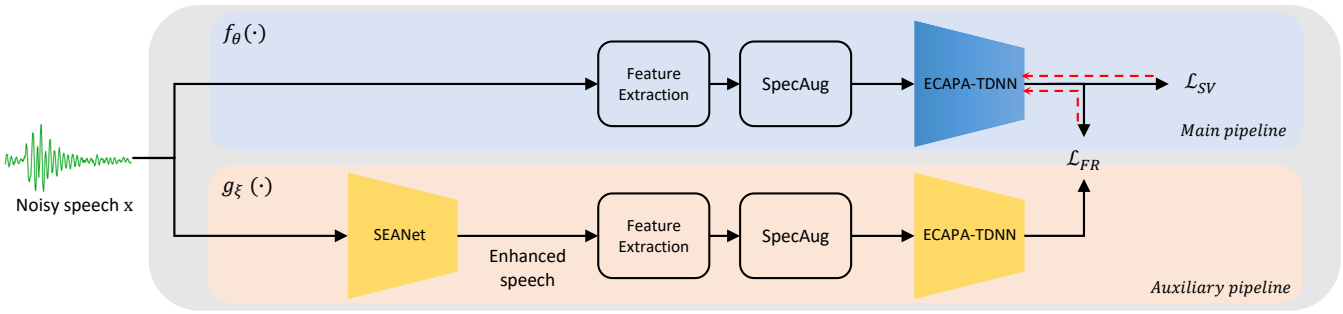

Fig. 1. *Overall structure of our proposed system for noise robustness. At the end of training, the pre-trained system $g_\xi(\cdot)$ is discarded, and the speaker classifier $f_\theta(\cdot)$ is only used for evaluations.*

work, which employ SEANet [5] and ECAPA-TDNN [6], respectively. To mitigate the adverse effect in noisy conditions, we train the combined SE and SV system $g_\xi(\cdot)$ with joint training. The output of the enhancement module, which is the enhanced speech signals, is fed into the back-end speaker classifier. During the training of the main pipeline, we freeze the parameters of the pre-trained model and use it only for extracting noise-robust speaker embedding.

Secondly, a single speaker classifier $f_\theta(\cdot)$ with a set of parameters θ, is trained to optimize the SV loss in an end-to-end manner. From the noisy speech signal $\mathbf{x}$, the single speaker classifier outputs embeddings $f_\theta(\mathbf{x})$, and the pre-trained system outputs embeddings $g_\xi(\mathbf{x})$. To enhance the noise robustness of the SV system, we introduce a feature-robust loss that minimizes the distance in the latent embedding space between both embeddings. Here, we define the mean squared error (MSE) between the l_2-normalized embeddings in two paths as the feature-robust loss:

$$\mathcal{L}_{FR}(\mathbf{x};\theta,\xi) = \left\| \frac{f_\theta(\mathbf{x})}{\|f_\theta(\mathbf{x})\|_2} - \frac{g_\xi(\mathbf{x})}{\|g_\xi(\mathbf{x})\|_2} \right\|_2^2 \quad (1)$$

Lastly, our proposed system is trained to discriminate the speaker-specific characteristic and alleviate the distortion of speech information by jointly optimizing the final objective as follows:

$$\mathcal{L}(\mathbf{x};\theta) = \mathcal{L}_{SV}(\mathbf{x};\theta) + \lambda \mathcal{L}_{FR}(\mathbf{x};\theta,\xi) \quad (2)$$

where λ is a weight coefficient to balance the speaker verification loss and the feature robust loss. During the training stage, we perform to optimize $\mathcal{L}$ with respect to θ only, not ξ. At the end of the training, we only keep the single SV systems $f_\theta(\cdot)$ which is then used for evaluations.

3. EXPERIMENTAL SETUP

3.1. Datasets

Experiments were conducted on the VoxCeleb1 [7] datasets. The development set contained 148,642 utterances from 1,211 speakers, and the test sets consist of 4,874 utterances from 40 speakers. The performance was evaluated with 37,720 enrollment-verification trial pairs. In training, we used the original VoxCeleb1 training dataset D along with the noise dataset D^N, which was synthesized using the MUSAN corpus [8] with a SNR randomly selected between 0 and 20 and the simulated room impulse response (RIR). For the evaluation under noise conditions, we used the VoxCeleb1 testset and synthesized the test data for each noise type with SNRs in the set $\{0, 5, 10, 15, 20\}$ using the MUSAN corpus.

3.2. Implementation details

For training the joint framework, we randomly cropped 2-second chunks from each utterance. The 80-dimensional log mel-filterbank energies with a 25ms window and a 10ms frame shift were used as input for SV systems. In the all experiments, we used the Adam optimizer with a cyclical learning rate of 0.001. The objective function for SV was AAM-Softmax [9] with a margin of 0.2 and a scale of 30. We set the weight factor λ to 1. Both the pre-trained cascade system and a single SV system were trained for 300 epochs. The performances were measured in terms of equal error rate (EER). We set the channel size C to 512 for ECAPA-TDNN. For SEANet as an enhancement module, we only used the generator and set the channel size of the initial convolution layer to 16.

4. RESULTS

In Table 1, we carried out the ablation studies for noisy conditions on the VoxCeleb1 testset. The evaluation was repeated 10 times, and we report the mean and standard deviation of the experiments. We observed that the performance of the ECAPA-TDNN trained only with clean dataset D drops sharply in noisy conditions. When compared with the mainstream approaches to enhance the noise robustness, in the case of a single SV system, the data augmentation strategies greatly improved the performance in noisy conditions and

Table 1. *Speaker verification performances obtained on VoxCeleb1 testset. We report the evaluation results for the mean and standard deviation of the 10 repeated experiments. Here, we use the jointly trained SEANet and ECAPA-TDNN to extract the embeddings in the process of training.*

Training dataset		Original (D)	Original and noise augmentation ($D + D^N$)		
Noise type	SNR	ECAPA-TDNN	ECAPA-TDNN	(Joint) SEANet+ ECAPA-TDNN	Proposed System
Original test set		3.50	2.55	2.63	**2.54**
Babble	0	23.36 ± 0.20	7.42 ± 0.27	$\mathbf{5.77 \pm 0.14}$	7.08 ± 0.23
	5	11.84 ± 0.11	4.29 ± 0.12	$\mathbf{4.13 \pm 0.08}$	4.26 ± 0.09
	10	6.42 ± 0.10	3.35 ± 0.05	3.46 ± 0.05	$\mathbf{3.32 \pm 0.06}$
	15	4.55 ± 0.08	2.93 ± 0.06	3.03 ± 0.06	$\mathbf{2.90 \pm 0.02}$
	20	3.82 ± 0.06	$\mathbf{2.67 \pm 0.04}$	2.84 ± 0.05	2.72 ± 0.05
Music	0	21.13 ± 0.18	7.32 ± 0.09	$\mathbf{6.80 \pm 0.11}$	7.21 ± 0.10
	5	11.23 ± 0.16	4.68 ± 0.13	4.61 ± 0.08	$\mathbf{4.53 \pm 0.08}$
	10	6.66 ± 0.13	3.62 ± 0.06	3.65 ± 0.08	$\mathbf{3.49 \pm 0.07}$
	15	4.65 ± 0.10	3.02 ± 0.07	3.14 ± 0.06	$\mathbf{3.00 \pm 0.04}$
	20	3.91 ± 0.05	2.76 ± 0.04	2.85 ± 0.05	$\mathbf{2.74 \pm 0.04}$
Noise	0	21.27 ± 0.33	7.35 ± 0.10	7.32 ± 0.07	$\mathbf{7.23 \pm 0.13}$
	5	12.97 ± 0.20	5.06 ± 0.10	5.01 ± 0.07	$\mathbf{4.87 \pm 0.09}$
	10	8.30 ± 0.11	3.93 ± 0.08	3.95 ± 0.08	$\mathbf{3.81 \pm 0.05}$
	15	5.89 ± 0.14	3.29 ± 0.06	3.32 ± 0.05	$\mathbf{3.23 \pm 0.07}$
	20	4.67 ± 0.10	2.95 ± 0.05	3.00 ± 0.03	$\mathbf{2.90 \pm 0.05}$
Average		10.04	4.30	**4.19**	4.22

also improved the clean scenarios. Similarly, the system that was jointly trained with the enhancement module showed significant performance improvement. In particular, we confirmed that the jointly trained system was effective at low SNR conditions. However, the performance improvement was slightly lower compared to the MCT-based single SV system because of the speech distortion generated during the de-noising process at high SNR conditions. Note that we used this system as the *pre-trained model* for extracting the embeddings, as mentioned in the previous section. As shown in the table, by utilizing the embeddings of pre-trained model combined with feature-robust loss, our proposed method can mitigate the effect of noise in low SNR conditions better than MCT-based single SV system and it can improve the performance in various SNR conditions. From the results, we demonstrate the possibility of developing robustness to noise using only a single SV model.

5. CONCLUSION

In this paper, we investigate a training method that aims to optimize a single noise-robust SV system. In order to learn useful information against noise from the pre-trained model, we introduce the feature-robust loss, which minimizes the distance in the latent embedding space. Experiments on the VoxCeleb1 dataset show that the proposed method improves the performance under various noise conditions and boosts the noise robustness even when using only a single SV system. Our future research will aim to explore the potential for applying this technology to blockchain systems and enhancing speaker authentication security in noisy conditions.

6. REFERENCES

[1] E. Variani, X. Lei, E. McDermott, IL. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2014, pp. 4052–4056.

[2] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2016, pp. 5115–5119.

[3] S. Han, Y. Ahan, K. Kang, and J. W. Shin, "Short-segment speaker verification using ecapa-tdnn with multi-resolution encoder," in *Proc. ICASSP*. IEEE, 2023.

[4] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[5] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," in *Proc. Interspeech*, 2020, pp. 1126–1130.

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

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

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

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF CVPR*, 2019, pp. 4690–4699.