2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

# Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation for Speaker Embedding Learning

Tomoka Wakamatsu* Sayaka Shiota* and Hitoshi Kiya*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: wakamatsu-tomoka@ed.tmu.ac.jp

***Abstract*—Data augmentation is essential for constructing reliable automatic speaker verification (ASV) systems. It is well known that data augmentation increases the number of utterances by adding noise and is effective in most methods. However, the number of speakers in the training data also plays an important role in enhancing ASV system performance. The robustness of speaker embedding networks, which are used in ASV systems, relies on the number of speakers present in the training data. To address this, we propose a method called pseudo-speaker augmentation, which utilizes a technique called vocal tract length (VTL) warping. By changing a parameter, the VTL warping technique alters speaker characteristics, allowing us to easily increase the number of speakers. Since the speaker embedding network aims to classify speakers, having a larger number of speakers enhances its robustness. In our experiments, the pseudo-speaker augmentation method improved the performance of the speaker embedding-based ASV system, achieving an equal error rate of 4.058% on the JTubeSpeech database.**

## I. Introduction

With the spread of spoken dialogue systems, automatic speaker verification (ASV), a biometric authentication technology using voice, has been actively researched in recent years. As many deep learning-based approaches advance, various deep neural network (DNN)-based ASV methods, such as x-vector [1] and ECAPA-TDNN [2], have been proposed and achieved remarkable performances. In these methods, a fixed-length feature vector called speaker embedding, extracted from the embedding layer of the DNN, is used to represent the personal characteristics of a speaker. In order to construct a reliable model in such a deep learning-based ASV system, a large amount of annotated data is required for model training. The most famous solution for insufficient annotated data is data augmentation by adding noise. Additionally, self-supervised learning (SSL) is also paid attention for ASV systems [3].

Data augmentation is an essential technique for constructing reliable speaker embedding networks. It is well known that data augmentation increases the number of utterances by adding noise and is effective in most methods [4], [5]. This technique also improves the generalization ability and robustness of the model by augmenting not only the amount of data but also the diversity of the data. In ASV tasks, data augmentation techniques to increase the number of utterances are generally used, such as the addition of noise or echoes [1], specaugment [5]. On the other hand, in a speaker embedding-based ASV system, it is known that the classification performance of the speaker embedding networks depends on the number of speakers in training data. Therefore, in constructing an ASV system, not only the number of utterances in the training data but also the number of speakers is considered an essential factor for improving the performance [6]. However, it is hard to collect large amounts of annotated data, including speaker labels. In [6], an augmentation method using speed perturbation [7] is proposed to increase the number of speakers at a small cost. It is reported that generating new speakers increases the range of speaker representation in the embedding space and improves the performance of ASV systems.

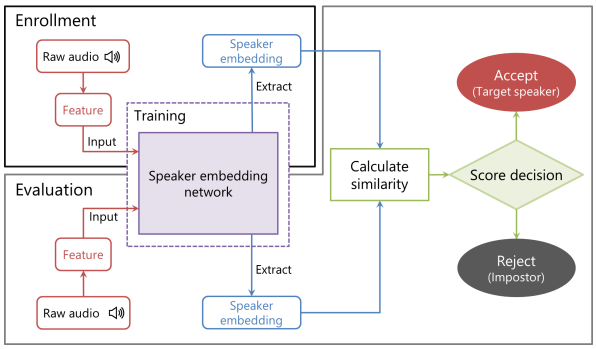


Fig. 1. Framework of speaker embedding-based ASV

Vocal tract length normalization (VTLN) [8] is used to counteract the effects of differences in vocal tract length. This technique can be reversely adapted to add variation to speech data. By changing a parameter, the vocal tract length warping technique alters speaker characteristics, allowing us to increase the number of speakers easily. Since the speaker embedding network aims to classify speakers, having a larger number of speakers enhances its robustness. Therefore, in this paper, we propose pseudo-speaker augmentation for speaker-embedding-based ASV by using the vocal tract length warping technique. In the experiments, we evaluated the effectiveness of the proposed pseudo-speaker augmentation with JTubeSpeech database [9]. The experimental results show that the best performance was obtained in the system with data augmentation for both the number of utterances and the number of speakers, and the effectiveness of speaker augmentation was confirmed.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

The rest of the paper is organized as follows. Section 2 describes a general speaker embedding-based ASV system. Section 3 will show the conventional utterance number augmentation by adding noise and the proposed pseudo-speaker augmentation with VTLP. Section 4 gives the detailed experimental setup and results, and the paper is concluded in Section 5.

## II. Speaker Embedding-based Automatic Speaker Verification

ASV is a binary classification task that determines whether the given voice is a registered speaker or not. Fig. 1 shows the framework of a speaker embedding-based ASV system, which has become mainstream in recent years. The system is constructed in the following flow:

First, a speaker embedding network is trained using a large amount of labeled data to extract features known as speaker embeddings, which represent the speaker's characteristics. The speaker embedding network is a neural network model trained for speaker classification and used as a feature extractor. The speaker embeddings are extracted from its embedding layer. Next, in the enrollment part, enrolled utterances are converted into acoustic feature vectors such as MFCC. Then, the features are fed into the trained speaker embedding network, and speaker embeddings are extracted. In the evaluation part, the speaker embeddings of test utterances are also extracted in the same manner. Subsequently, the similarity scores between the extracted speaker embedding vectors are computed, and the scores are used for determining the decision whether to accept or reject.

## III. Pseudo-speaker Augmentation

### A. Data augmentation using simulated noises

In most ASV systems in real environments, test utterances contain background noise. To construct a robust ASV system, it is necessary to include training data similar to the test environment. Also, a variety of training data is essential to build a more robust system that can adapt to different or unknown environments. However, it is very costly that collect large amounts of speech data recorded in many kinds of different environments. Data augmentation techniques are widely used in various tasks to increase simulated variation in training data and to improve the robustness of the trained model. Adding noise or background sounds to original data is a common data augmentation method Incorporating different types and levels of noise into the training data allows for better performance in real-world scenarios with varying acoustic conditions.

### B. Pseudo-speaker augmentation using Vocal Tract Length Warping

There are various factors that cause different speakers' voices to have different acoustic features, one of which is called vocal tract length (VTL). If the vocal tract is modeled as a uniform tube, it is possible to scale the frequency axis of the spectra of two speakers with different vocal tract lengths using a warping factor. It is known that such differences

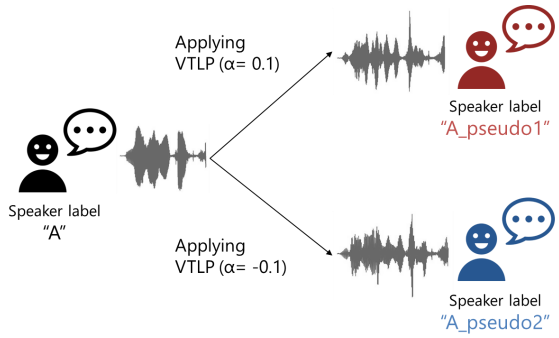


Fig. 2. Flow of making pseudo-speaker by VTLP

in vocal tract length affect recognition accuracy in speech recognition, and vocal tract length normalization (VTLN) [8] is used as a method to counteract this effect. VTLN removes inter-speaker variation by linearly warping the frequency axis of each speaker's spectrogram using the optimal warp factor for each speaker. A reverse adaptation of the VTLN method is the vocal tract length perturbation (VTLP) [10] approach, which adds variation to the input data by adding variation. In this paper, these frequency warping techniques are used as a data augmentation method to increase the number of speakers by attaching a different speaker ID to the processed speech than the original voice, like Fig. 2. If $\omega$ is the normalized frequency of the original voice and $\omega$' is the frequency after augmentation and contraction, it can be expressed as:

$$\omega' = \omega + 2\arctan\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)}, \tag{1}$$

where $\alpha$ is the VTL warping factor.

## IV. Experiment

We conducted ASV experiments to show the effectiveness of the proposed pseudo-speaker augmentation using VTLP.

### A. Database

JTubeSpeech [9] was used to train and evaluate the ASV systems. This is a database of mainly spoken Japanese speakers extracted from videos uploaded to YouTube. It is noted that since the videos are automatically collected, JTubeSpeech also includes other languages, such as English and Chinese. A dataset of JTubeSpeech is designed for ASV with a large number of speakers. In the experiments, we used the single-speaker dataset of the JTubeSpeech subset as the training dataset. The training dataset contains 502 hours of utterances from 1,795 speakers, and the development dataset contains 138 hours of utterances from 1,795 speakers. A total of 640 hours of utterances were used as the training dataset. The test dataset contained 20,976 trials by 92 speakers. The evaluation trials consist of 228 pairs with target speakers and 20,748 pairs with impostors. The speakers in the test set are all manually annotated as Japanese speakers.

MUSAN [11] was used for utterance augmentation. It contains 42 hours of music of various genres, 60 hours of

TABLE I
ARCHITECTURE OF CNN USED AS SPEAKER EMBEDDING EXTRACTOR

| | Layer | Kernel size | Input x Output |
|---|---|---|---|
| 1 | Conv1d | 5 | 40 x 128 |
| 2 | Conv1d | 3 | 128 x 128 |
| 3 | Conv1d | 3 | 128 x 128 |
| 4 | Conv1d | 3 | 128 x 64 |
| 5 | Fc | - | (64x3) x 512 |
| 6 | Fc | - | 512 x (*n_classes*) |

conversations in 12 languages, and more than 900 noise types. Noise type which contains approximately 6 hours of various types of noise, including mechanical, non-mechanical, and environmental sounds.

### B. Experimental setup

The architecture of the convolutional neural network (CNN) used as the speaker embedding extractor is shown in Table I. The output vectors of the first fully connected layer in Table I were regarded as the speaker embedding vectors. The input features of CNN were a 40-dimensional MFCC with a frame length of 25 ms. The network was trained for 100 epochs with the Adagrad optimizer [12] and Cross-Entropy Loss.

For the JTubeSpeech training dataset, we applied both data augmentation increasing the number of utterances by adding noise and increasing the number of speakers by VTLP described in section 3. In this paper, we refer the noise-adding data augmentation as utterance augmentation. The comparison methods were shown as follows:

(A). None
No data augmentation was applied to the training data. This was referenced as the baseline system.

(B). Noise
For utterance augmentation, noise data was randomly selected from the MUSAN noise data set, and the signal-to-noise ratio (SNR) was randomly set from [-5, 0, 5, 10, 15]. We applied utterance augmentation by using different noises and SNRs for each utterance. In this condition, the number of utterances was augmented by a factor of three by adding two types of noise data.

(C). Noise (large)
The utterance augmentation was applied in the same manner as condition (B). To compare the performance of a larger amount of data expansion only by adding noise, the SNRs were set to all of [-5, 0, 5, 10, 15], and the number of utterances was augmented by a factor of six.

(D). VTLP (all)
The pseudo-speaker augmentation was applied to all speakers in the training data, and the number of speakers was augmented by a factor of three. The warping factor $\alpha$ for VTLP was set to -0.1 or +0.1 for all utterances.

(E). VTLP (select)
Pseudo-speaker augmentation was performed in the same manner as condition (D). Figure 3 shows the distribution of the degree of changing speaker characteristics before and after adapting VTLP. Therefore, we decided to select the pseudo-speakers whom can use as different speakers from the original one adequately. Finally, the number of speakers is about twice as large, which is a smaller number than in condition (D).

(F). VTLP (select) + Noise
To investigate the effectiveness of combining both utterance number augmentation and speaker number augmentation, almost the same number of utterance augmentation was applied to condition (E) as in condition (B).

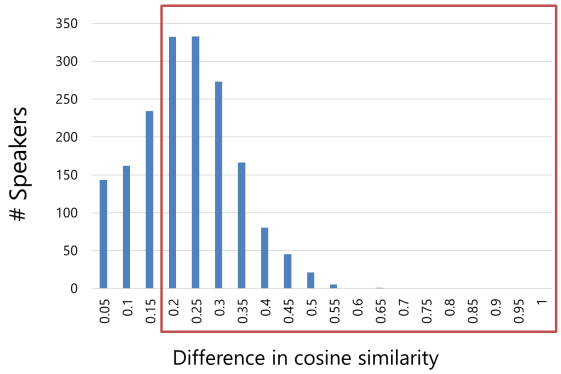


Fig. 3. Distribution of variation in speakability by VTLP

In creating pseudo-speakers, it is important that the individual characteristics after VTLP differ significantly from the original speaker. Therefore, we prepared condition (D) to test whether the degree of change in personal characteristics due to VTLP affects the accuracy of the ASV system. Specifically, we extracted the speaker embeddings of both the pseudo-speaker and the original speaker using one of our trained speaker classification models, and computed the average of the difference in the cosine similarity between them for each speaker. The average of the differences in similarity per speaker for a single warping factor is shown in Fig. 3. This confirms that the change in personal characteristics at a given value of the warping factor varies from speaker to speaker. For some speakers, however, the warping coefficients set in this study were not sufficient to treat the pseudo-speaker as a different speaker. Therefore, we used as VTLP (select) only pseudo-speakers with a difference in cosine similarity greater than 0.2. By implementing these conditions, we examined the usefulness of each data augmentation method in terms of the number of speakers and utterances. The amount of data in each condition is shown in the table II.

### C. Evaluation metrics

To evaluate the comparison methods, we used two kinds of test trials provided in the JTubeSpeech dataset. One trial is defined as a core task, whose segments are set to under

TABLE II
DATA AMOUNT OF THE TRAINING DATASET IN EACH CONDITION

| | Augmentation | # Speakers | Total hours |
|---|---|---|---|
| (A) | None | 1,795 | 502.49 |
| (B) | Noise | 1,795 | 1,510.17 |
| (C) | Noise (large) | 1,795 | 3,014.90 |
| (D) | VTLP (all) | 5,385 | 1,506.32 |
| (E) | VTLP (select) | 3,681 | 957.59 |
| (F) | VTLP (select) + Noise | 3,681 | 2,875.49 |

TABLE III
EERS (%) FOR EACH AUGMENTATION METHOD.

| | Augmentation | EER (%) | |
|---|---|---|---|
| | | Core | Short |
| (A) | None | 6.140 | 7.456 |
| (B) | Noise | 5.104 | 8.333 |
| (C) | Noise (large) | 5.104 | 8.333 |
| (D) | VTLP (all) | 7.018 | 9.211 |
| (E) | VTLP (select) | 6.579 | 7.822 |
| (F) | VTLP (select) + Noise | **4.058** | **6.777** |

five seconds, and another one is referenced as a short task, whose segments are set to under two seconds. These trials were released as the evaluation tasks of JTubeSpeech-ASV[1]. The evaluation metric was the equal error rate (EER) based on a comparison of cosine similarity between the speaker embeddings.

### *D. Experimental results*

Table III shows the EERs under the core and short trials for each comparison condition. First, we discuss the results under the core trials. Comparing the results of (A) with (B) and (C) indicates that (B) and (C) improved performance compared to the baseline, attributed to the inclusion of utterance augmentation. On the other hand, when comparing (B) with (C), since only utterance augmentation was performed, the EERs were the same. Since both augmented data were generated by adding noise, the data diversity was limited, and it seems to lead to the saturated result. Comparing (D) with (E), both using the pseudo-speaker augmentation, the EER of (E) is 0.439 points better than that of (D). In both (D) and (E), pseudo-speaker augmentation was applied, and in (E), speakers were selected for which the change in their characteristics due to VTLP was significant. This indicates that when the speakers were chosen to apply VTLP, as in (E), VTLP can properly function as a data augmentation method. However, both (D) and (E) perform worse than the baseline. One reason for this result is that whereas the diversity of the speakers was increased, the number of utterance data for each speaker was not enough. (F) achieved the highest performance among all conditions. Especially comparing (F) with (C), while the EER of (C) seems saturated, the EER of (F) was 1.046 points better than that of (C) with almost the same amount of data. It denotes that the combination of utterance augmentation and pseudo-speaker augmentation contributes to expanding the data diversity, and it helps to improve performance. The short trials are challenging tasks, as stated in the Short-duration Speaker Verification Challenge(SdSVC) 2021[13]. In the results under these short trials, the performances were getting lower than the performances of the core trial. Comparing the results of (A) with those from (B) to (E), it appears that data augmentation is not effective in improving performance. However, comparing (E) with (B) and (C), the performance with the pseudo-speaker augmentation by VTLP(select) is better than the utterance augmentation. This trend is in contrast to the core trials. For short utterance trials, the speaker embedding quality is more important to represent speaker features accurately. The results suggest that the augmentation of the number of speakers may be effective in spreading the representation space of the speaker embedding. Consequently, the proposed pseudo-speaker augmentation was found to be effective when combined with the utterance augmentation in both the core and short trials.

[1]https://github.com/sarulab-speech/jtubespeech

## V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed pseudo-speaker augmentation by using VTLP for robust ASV systems. Data augmentation by adding noise is basically used for the state-of-the-art speaker embedding models. On the other hand, the larger the number of enrollment speakers for the speaker embedding models, the more robust the performances. Therefore, the pseudo-speaker augmentation was performed by using VTLP. Additionally, we described that the pseudo-speaker selection was effective for the ASV systems. The experimental results showed that the best performance was obtained when both the pseudo-speaker augmentation and the utterance augmentation were applied simultaneously. In the future, we will investigate the effectiveness of the proposed method on state-of-the-art systems such as x-vector and ECAPA-TDNN.


## ACKNOWLEDGMENT

This work was supported in part by SECOM Science and Technology Foundation.


## REFERENCES


[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Proc. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, 2018.

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

[3] Z. Chen, S. Chen, Y. Wu, *et al.*, "Large-scale self-supervised speech representation learning for automatic speaker verification," *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, 2022.

[4] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, "Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification," *Proc. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 205–209, 2018.

[5] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[6] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," *Proc. Interspeech 2019*, pp. 406–410, 2019.

[7] T.Ko, V.Peddinti, D.Povey, and S.Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[8] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on speech and audio processing*, vol. 6, no. 1, pp. 49–60, 1998.

[9] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "Jtubespeech: Corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[10] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, p. 21, 2013.

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

[12] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization.," *Journal of machine learning research*, vol. 12, no. 7, 2011.

[13] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2021: The challenge evaluation plan," *arXiv preprint arXiv:1912.06311*, 2019.