Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation Considering Speaker Variability for Speaker Verification

Hengyi Zou* Sayaka Shiota*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: zou-hengyi@ed.tmu.ac.jp

***Abstract*—In this paper, we propose an advanced method for pseudo-speaker augmentation using vocal tract length perturbation (VTLP) for automatic speaker verification (ASV) systems. The state-of-the-art ASV systems based on speaker embeddings require a substantial amount of training data to construct a reliable speaker embedding extractor. Traditional data augmentation methods for ASV typically focus on increasing the corpus size, while ensuring sufficient diversity of distinct speakers is also crucial for improving accuracy. A previous study has reported that VTLP is used as an effective pseudo-speaker generation method, and increasing the number of speakers through VTLP can enhance ASV performance. However, the previous method has demonstrated limitations in the number of pseudo-speakers that can be effectively used, indicating that these methods may not be sufficiently effective. Therefore, this paper proposes increasing the number of pseudo-speakers available for data augmentation by setting the VTLP parameters to ensure diversity for each speaker. The experimental results show that the proposed pseudo-speaker augmentation method can significantly improve the performance of ASV system based on ECAPA-TDNN.**

I. Introduction

In recent years, people's concern for security has been increasing as various systems have been digitized, and research on biometric authentication techniques has become increasingly important [1]. Biometric information used for authentication includes fingerprints, iris, veins, etc. Among them, biometric technology using voice is automatic speaker verification (ASV). ASV is expected to have practical value as a biometric authentication technology because of its low cost of implementation and the growing demand for online biometric authentication.

State-of-the-art (SoTA) ASV systems are based on speaker embeddings using deep learning, represented by x-vector [2] and ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network) [3]. These SoTA ASV systems employ deep learning techniques for extracting speaker embeddings, leveraging advanced methods to achieve efficient and accurate results. It is well known that in order to extract more expressive speaker embeddings, deep neural networks (DNNs) need to be trained on a large amount of training data. In addition to the existing training data, data augmentation methods are used to further enhance the training data and their effectiveness in speaker recognition has been proven [2]–[7].

In conventional ASV systems, the method widely used for data augmentation is to increase the number of utterances per speaker by adding noise. Futhermore, speed perturbation is also employed to further augment the data by varying the speaking speed of the utterances, thereby increasing the variability and robustness of the training dataset [8]. Additionally, the other research has reported that sufficient diversity among distinct speakers is crucial [9], [10]. To address this, vocal tract length perturbation (VTLP) [11], [12], which is derived from vocal tract length normalization (VTLN) [13], [14], has been proposed as a method to enhance speaker variability. By perturbing the frequency axis of voice, it is possible to generate a pseudo-speaker and increase the total number of speakers by adding it to the original training data, therefore it could increase the complexity of speaker embedding. The previous study [9] has reported that ASV performance is improved by selecting only pseudo-speakers with sufficiently large speaker variability, rather than simply adding pseudo-speakers. Nevertheless, this approach reduces the number of pseudo-speakers available for augmentation, thereby not fully achieving the original goal of increasing the number of speakers. To address this limitation, a refined method is proposed that balances the variability of speaker characteristics with the quantity of pseudo-speakers.

In this paper, we build on the prior work of [9] by proposing a method to further increase the number of pseudo-speakers by adjusting VTLP parameters to ensure sufficient variation in speaker characteristics. Our proposed method involves calculating the cosine similarity of speaker embeddings before and after applying VTLP for each speaker. For speakers whose characteristics do not exhibit significant variation, we modify the VTLP parameters to achieve the necessary changes. Speakers who exhibit sufficient variation are then added to the speaker pool as pseudo-speakers. This iterative adjustment process ensures that the generated pseudo-speakers introduce sufficient variability, thereby enhancing the effectiveness of our speaker verification system. In the experiments, an ASV system using ECAPA-TDNN is constructed, and evaluated it's performance when various data augmentation were applied to the training data. The experimental results confirm the significance of incorporating a diverse range of speaker characteristics into the training data. The system's performance

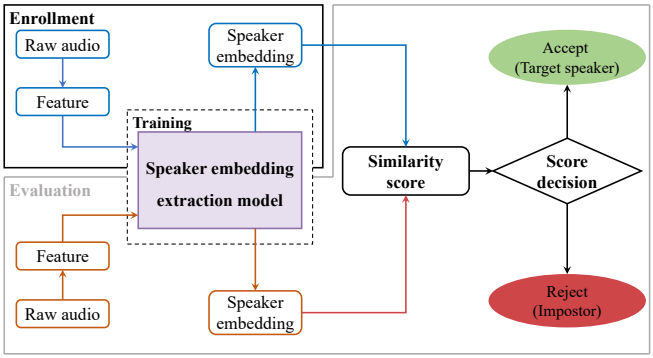

Fig. 1: Flow of speaker embedding-based ASV

improved most notably with the data augmentation method that maximized the dataset size while accounting for speaker variability. This was achieved by increasing the number of utterances per speaker through adding noise and expanding the number of speakers by VTLP. The importance of including a variety of speaker characteristic in the training data was confirmed.

The rest of the paper is organized as follows. Section 2 describes a genral ASV system based in speaker embedding. Section 3 introduced 2 kinds of augmentation. Section 4 will show how to select pseudo-speaker by variability of speaker characteristic. Section 5 gives details of experiment and results, and section 6 concludes this paper.

II. Speaker verification based on speaker embedding

Speaker recognition can be classified into the following two categories. The first is speaker identification in a multi-level classification task that identifies the most likely speaker from several registered speakers. Another is ASV in the binary classification task, which determines whether the input speech belongs to the registered speaker himself or not. Figure 1 illustrates the flow of the ASV system based on a recent speaker embedding technique. The ASV system comprises enrollment, evaluation, and training units for the speaker embedding extraction model, utilized in both enrollment and evaluation. Initially, the speaker embedding extraction model is trained using a large dataset to recognize a massive number of different speakers. The intermediate output of this model, which can accurately identify speakers, is used as the speaker embedding for both enrollment and evaluation. In the enrollment phase, the utterance of the enrolled speaker is converted into features, which are input to the model to extract the speaker embedding. The same procedure applies to the evaluation phase for the test speaker. The similarity between the extracted embeddings is calculated using cosine similarity or other metrics, and compared with a threshold to determine speaker identity.

III. Data augmentation for speaker verification

To enhance the performance of the speaker embedding extraction model, a substantial amount of training data is required. Besides using existing databases, data augmentation is widely employed to simulate additional training data. In ASV, data augmentation typically involves increasing the number of utterances through methods such as adding noise and music. Moreover, augmenting the number of speakers using techniques like VTLP has been shown to improve accuracy [9], [15].

A. Speech augmentation using simulated noises

This method expands the number of utterances by adding noise and music to speech data. Meanwhile, it's often used in deep neural network training, helps stabilize training and improve robustness by simulating a noisy environment [16]. Augmenting the number of utterances in ASV can stabilize the training of embedding and extraction models for speaker identification.

B. Pseudo-speaker generation by VTLP

VTLN is a technique from speech identification to normalize the acoustic feature variability across different speakers. In this method, pseudo-speakers are generated by transforming the frequency axis of the logarithmic amplitude spectrum of speech. The normalization frequency of the original voice is ω, frequency after perturbation is ω', if the frequency expansion coefficient is α, it is represented by the formula (1).

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

In addition, the level of pitch change could be adjusted by changing the parameter of the frequency expansion coefficient. In speech recognition, VTLN is used to normalize the vocal tract length to remove the effect of speaker characteristic, however in our research, this formula is used as VTLP to add variation to the vocal tract length. So far, it has been reported that VTLP can be used for data augmentation in ASV by adding vocal length perturbation [11], [12]. In these methods, the number of speakers is augmented by using the voice of a new speaker, which is processed to a different pitch from the original voice by applying VTLP to the training data, to increase the number of speakers in a pseudo form.

IV. Pseudo-Speaker generation considering speaker variability

A. Speaker verification by pseudo-speaker generation using VTLP

In ASV using data augmentation with VTLP described in the previous chapter, two methods were proposed: one consistently applies a frequency expansion coefficient to increase speakers, while the other selects pseudo-speakers with significant variability in speaker characteristics. The first method has the advantage of simply increasing the number of speakers. However, incorporating pseudo-speakers with small variability in speaker characteristics could lead to problem, as it can confuse the network and make it difficult to distinguish between the original speaker and the pseudo-speaker. On the other hand, the method that selects only pseudo-speakers with large speaker

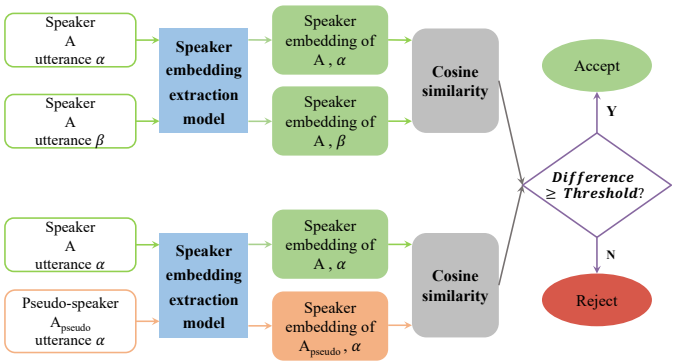

Fig. 2: Flow of calculating the amount of speaker variability

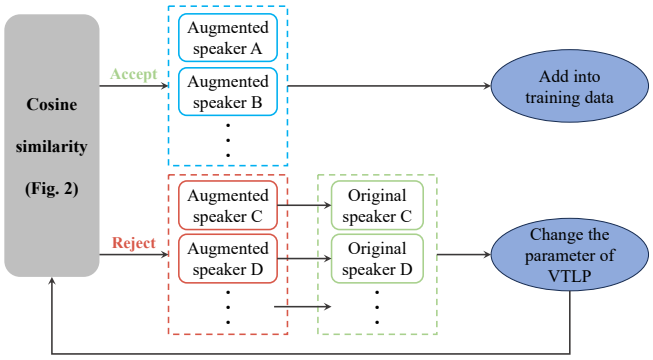

Fig. 3: Flow of Parameter adjustment considering of speaker variability

variability has better performance in ASV. This effect is likely due to the difficulty the network faces when distinguishing between the original speaker and a pseudo-speaker with small variability in characteristics. When the network assumes that a speaker with small variability is distinct from the original speaker, making accurate distinctions becomes challenging, which reduces the network's accuracy. Selecting only voices with substantial variability in speaker characteristics as pseudo-speaker voices can stabilize the training of the speaker embedding extraction network by excluding pseudo-speakers with minimal variability. Nevertheless, this approach also reduces the number of pseudo-speakers available for inclusion in the training data, which counteracts the primary objective of data augmentation. Consequently, the number of pseudo-speakers that can be added to the training data is reduced, and the cost is that the number of speakers cannot be increased sufficiently for the main purpose of data augmentation.

B. Data selection by variability of speaker characteristic

Pseudo-speaker generation involves applying VTLP to all speech data and assigning new speaker labels to the processed speech. The variability of speaker characteristics is calculated using cosine similarity between original and pseudo-speaker embeddings. Only those pseudo-speakers with variability above a certain threshold are added to the training data. Figure 2 outlines the process of calculating speaker variability. The cosine similarity is computed between the speaker embeddings of Speaker A's utterances α and β, and between the embeddings of Speaker A's utterance α and the pseudo-speaker A_{pseudo}'s utterance α. The difference in cosine similarity is then assessed by comparing the embeddings of Speaker A's utterances with each other and those of Speaker A's utterance with the pseudo-speaker A_{pseudo}'s utterance. This difference represents the variability of speaker characteristics. Since the cosine similarity between Speaker A's utterances can vary, multiple utterances from Speaker A will be used as reference utterances. Multiple cosine similarities will be calculated and averaged to determine the overall cosine similarity between Speaker A's utterances. In contrast, to obtain the similarity between the pseudo-speaker A_{pseudo}'s speech and Speaker A's speech, the cosine similarity is calculated between the embeddings of Speaker A's speech and all pseudo-speaker A_{pseudo}'s speech generated by applying VTLP to Speaker A's speech. If the calculated variability of speaker characteristics exceeds a predefined threshold, the pseudo-speaker A_{pseudo}'s speech is added to the training data. If it is below the threshold, it is not used as the pseudo-speaker.

C. Pseudo-speaker selection by variability of speaker characteristic

This paper proposes generating new pseudo-speakers with modified VTLP parameters for those initially deemed to have low variability. A previous study reported that the number of pseudo-speakers was reduced to about 40% through data selection to improve ASV performance. To address this limitation, we propose increasing the data for speaker augmentation by applying VTLP with modified parameters to the original speakers of excluded pseudo-speakers. Figure 3 provides an overview of the proposed method. Initially, the procedure in Fig. 2 is used to assess the variability of pseudo-speakers. For those with low variability, VTLP with modified parameters is applied to generate new pseudo-speakers with increased variability. This enhanced variability improves the effectiveness of data augmentation. The process in Fig. 2 is repeated for these new pseudo-speakers to determine their inclusion in the training data. By adjusting parameter variations and iteratively applying the method shown in Fig. 3, a greater number of pseudo-speakers can be automatically selected. However, as the absolute value of the VTLP parameter increases, speech distortion also increases. When this distortion reaches a threshold beyond which the speech can no longer be recognized as intelligible, the proposed data augmentation method will be halted.

V. EXPERIMENT

In order to verify the effectiveness of the proposed method, data augmentation for the number of speeches and speakers was applied under various conditions, and ASV experiments were conducted.

TABLE I: Number of speakers and total duration of speech for each data augmentation condition

Augmentation condition	Num of speakers	Data duration (hrs)	Average data duration per speaker(hrs)
(A) No augmentation	1792	498	0.28
(B) Noise	1792	1495	0.83
(C) VTLP (All)	5376	1494	0.83
(D) VTLP (Select)	2868	794	0.44
(E) VTLP (Proposed)	4086	1128	0.63
(F) VTLP (Proposed+Noise)	4086	5380	3.00

A. Database

JTubeSpeech-ASV [17] was used to train and evaluate the ASV system, it is a speech corpus consisting of 900 hours of speech data automatically collected from YouTube videos. It consists of audio data of only one speaker appearing in one video, and one channel is considered as one speaker. This dataset mainly includes Japanese audio, and it also includes languages such as English, Chinese, and Korean. Among the subset for ASV of JTubeSpeech-ASV used for learning and evaluating the speaker embedding extraction model, the training dataset consists of 107,271 utterances from 1792 speakers. The test dataset for ASV is also from the JTubeSpeech-ASV dataset. The test dataset consists of 20,976 speech utterances from 92 speakers and amounting to 20,976 trials. Of the 20,976 sets, 228 are trials within the same speaker, and 20,748 are trials between different speakers. The sampling frequency is 16 kHz. The MUSAN database [18] is used for adding noise and music. The MUSAN database contains 42 hours of music of various genres, 60 hours of conversations in 12 languages, and over 900 types of noise. In the experiments, only the noise subset of the MUSAN database is used. The subset that this paper used contains a total of about 6 hours of noise, including mechanical and environmental noises.

B. Experimental setup

In the experiments, an ASV system based on speaker embedding is constructed. The ECAPA-TDNN is used as the speaker embedding extraction model, and it is known to have higher performance. A logarithmic mel filter bank was used as the input features of the model. The speaker embedding was extracted as a 512-dimensional vector using the output of the middle layer of the ECAPA-TDNN.

To verify the relationship between ASV performance and speaker characteristic variability based on different parameter settings, three methods were compared: the conventional method applying VTLP with a single parameter to all speakers (All), the conventional method selecting augmented speakers based on characteristic variability (Select), and the proposed method applying VTLP with various parameters to the speaker characteristics. In the experiments, three proposals have been prepared for augmenting speakers. The comparison conditions are shown below, and the number of data sets for training in each condition is summarized in Table 1.

(A) No augmentation

Only the training dataset of JtubeSpeech-ASV, no data augmentation was applied.

(B) Noise

For the training dataset of JtubeSpeech-ASV, the number of utterances is augmented by noise superimposed. The noise data was randomly selected from MUSAN's noise dataset, and the SNR was set to 0. Two types of noise data were superimposed for each voice data, and the number of utterances per speaker was tripled.

(C) VTLP (All)

Augmenting the number of speakers by VTLP on the JtubeSpeech-ASV training dataset. The VTLP (All) applies VTLP to 1,792 speakers in the training dataset. In the conventional method (All), VTLP is applied to the 1,792 speakers in the training dataset. The frequency expansion coefficients are set to 0.1 and -0.1, and VTLP is applied to each voice to generate two voices each, and 3,584 pseudo-speakers were added by assigning a new speaker label to the speech as if it were spoken by a different speaker from the original speaker.

(D) VTLP (Select)

Among the pseudo-speakers generated by the VTLP (All), by using ECAPA-TDNN as the extraction model for speaker embedding, only those pseudo-speakers with a large variability of speaker characteristics from the original speaker are selected for augmenting the number of speakers.

(E) VTLP (Proposed)

Among the pseudo-speakers generated by the VTLP (All), by using ECAPA-TDNN as the extraction model for speaker embedding, it is applied to the pseudo-speakers which are excluded by condition D, therefore the amount of data could be increased. The frequency scaling factor will be two cases: increasing in increments of 0.01 from 0.1 to 0.17 or decreasing in increments of 0.01 from -0.1 to -0.17.

(F) VTLP (Proposed + Noise)

The experiment augmented the data for both the number of augmented speakers and the number of utterances by adding noise and music to the speech in condition (E). Noise data were randomly selected from the MUSAN noise data set, and SNR was set to 0. Two types of noise data were superimposed on each speech data, and the number of utterances per speaker was increased by a factor of three.

The experiment used the equivalent error rate (EER) as an evaluation index. The EER is calculated from the point where the acceptance rate of a person by another person is equivalent to his/her rejection rate, and the smaller the value is, the better the accuracy is evaluated.

C. Experimental Result

Table 2 shows the EERs of the ASV results for each of the conditions (A)-(F). First, comparing the two conditions (A) and (B), the EER for condition (B) is 0.781 points lower than that for condition (A), which confirms the effectiveness of speech augmentation by adding noise and music. Futhermore, comparing the conventional method (C) and (D), both EERs are almost the same as the EER of (A) without data augmentation. It is obvious that (C) has the highest EER overall, even though the

amount of data is almost the same as in (B). From the results above, it suggests that the ECAPA-TDNN does not show much improvement in accuracy compared to the conventional method. About comparing the three conditions (C)-(E), the EERs of (D) and (E) are lower than those of (C) because the variability of speaker characteristic is taken into account while the number of speakers is augmented using VTLP. The fact that the EERs for both cases improved despite the smaller number of speakers and data volume compared to (C) confirms the necessity of taking the variability of speaker characteristic into account. Among these three conditions (C)-(E), the proposed method (E) has the lowest EER, which is 0.945 points lower than that of (A). Compared to both conventional methods, the proposed method (E) shows a significant improvement in accuracy, and its EER is 0.939 points lower than that of (D). In both (D) and (E), the number of speakers is augmented to account for the variability of speakerness, however (E) uses more pseudo-speakers for augmentation than (D). This confirms the effectiveness of using more pseudo-speakers by adjusting the parameters. In (F), the number of utterances per speaker is further augmented from (E), in which the number of speakers is augmented using the proposed method, and the EER is reduced by 0.651 points compared to (E). At the same time, the EER was 0.815 points lower than that of (B), which only expanded the number of utterances, resulting in the best performance among all conditions. From the above results, it is confirmed that, in data augmentation for ASV based on speaker embedding, the combined use of noise-superimposed utterance augmentation and speaker number augmentation with consideration of variability of speaker characteristic is very effective in improving the performance.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method for augmenting training data for ASV models by generating pseudo-speakers through VTLP and considering the variability of speaker characteristic for selection. Simultaneously, we augment the number of speakers by varying the VTLP parameters. In the experiment, data augmentation was applied to the JTubespeech-ASV training dataset by adding noise to increase the number of utterances, generating pseudo-speakers to increase the number of speakers, and using both methods. An ASV model based on speaker embeddings was trained and evaluated the performance of the ASV system. The experimental results demonstrated that the best performance was achieved when more pseudo-speakers were added to the training data, with a focus on selecting those with significant variations in speaker characteristics.

For future work, it is considered necessary to explore more quantitative methods for setting the most appropriate frequency scaling factors and thresholds. Additionally, since there can be differences in the variations of speaker characteristics even within the same speaker's speech, it is necessary to consider more suitable methods for selecting pseudo-speakers.

TABLE II: EER (%) of speaker verification for each condition

Augmentation condition	**ECAPA EER(%)**
(A) No augmentation	6.984
(B) Noise	6.203
(C) VTLP (All)	7.018
(D) VTLP (Select)	6.978
(E) VTLP (Proposed)	6.039
(F) VTLP (Proposed+Noise)	5.388

ACKONWLEDGMENT

This work was supported in part by JSPS KAKENHI (Grant Number JP24K14993), and ROIS DS-JOINT (022RP2024) to S. Shiota. We would like to thank Ms. Ono for her valuable assistance with the statistical analysis and her insightful suggestion on the research methodology.

REFERENCES

[1] Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019.

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

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[4] S. Wang, J. Rohdin, O. Plchot, L. Burget, K. Yu, and J. Černocký, "Investigation of specaugment for deep speaker embedding learning," in *Proc. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] P. S. Nidadavolu, V. A. Iglesias, J. Villalba, and N. Dehak, "Investigation on neural bandwidth extension of telephone speech for improved speaker recognition," in *Proc. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[6] C.-L. Huang, "Exploring effective data augmentation with tdnn-lstm neural network embedding for speaker recognition," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[7] Y. Zhu, T. Ko, and B. K.-W. Mak, "Mixup learning strategies for text-independent speaker verification," in *Proc. Interspeech*, 2019.

[8] H. Yamamoto, K. A. LEE, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Proc. Interspeech*, 2019.

[9] T. Wakamatsu, S. Shiota, and H. Kiya, "Vocal tract length perturbation-based pseudo-speaker augmentation for speaker embedding learning," in *Proc. 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023.

[10] K. Zhou, Q. Yang, X. Sun, and S. Liu, "A deep speaker embedding transfer method for speaker verification," in *Proc. Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 1*, Springer, 2020, pp. 369–376.

[11] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML workshop on deep learning for audio, speech and language*, vol. 117, 2013, p. 21.

[12] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996.

[13] K. Johnson, "Vocal tract length normalization," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[14] L. D. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 49–60, 1998.

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Proceedings of INTERSPEECH*, 2015, pp. 3586–3589.

[16] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, 2007.

[17] 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:2112.09323*, Dec. 2021.

[18] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.