Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Application of bandwidth extension with no learning to data augmentation for speaker verification

Haruna Miyamoto, Sayaka Shiota, Hitoshi Kiya

Tokyo Metropolitan University
Faculty of System Design, Department of Computer Science, Japan

Abstract

In this paper, we propose a data augmentation scheme with bandwidth extension (BWE) for deep neural network (DNN)-based automatic speaker verification (ASV) systems. One approach used for DNN-based ASV systems, named "x-vector," requires a large amount of training data. In particular, the performance of x-vector-based ASV systems is amongst the highest when using a large amount of wideband (WB) data. However, when the amount and variety of data are limited, it is important to use data augmentation schemes. If BWE methods can be used as data augmentation schemes for x-vector-based systems, the issue with the amount and variety of data would be relaxed. In some reports, the use of WB data extended from narrowband (NB) data by DNN-based BWE has already been considered. Recently, the authors reported on the effectiveness of BWE methods for machine learning frameworks. Additionally, the quality of speech generated by non-learning-based BWE is almost the same as that by learning-based BWE. Therefore, in this paper, we aim to demonstrate that several non-learning-based BWE methods are useful for data augmentation for x-vector-based ASV systems. The results of an experiment done using a speakers in the wild and NIST SRE databases showed that a system using the proposed scheme reduced error 22.7% compared with our baseline system.

1. Introduction

The development of state-of-the-art automatic speaker verification (ASV) systems is an active research area. For practical applications of ASV, ASV systems are required to be highly reliable, convenient, and low-cost. Thanks to the latest ASV approaches, the i-vector-based approach [1], deep neural network (DNN)-based approaches [2–4], and the probabilistic linear discriminant analysis (PLDA) classifier [5], the performance of ASV systems is high. In particular, the success of the x-vector-based approach [6] has significantly improved ASV performance. However, the x-vector-based approach requires a huge amount of training data to achieve high performance. When the amount and variety of data are limited, it is important to use data augmentation schemes. Therefore, data augmentation schemes for x-vector-based ASV systems have been reported [7–10].

The National Institute of Standards and Technology (NIST) has provided several narrowband (NB) databases for the ASV research area. When NB signals can be used with wideband (WB) signals as training data for x-vector-based ASV systems, the performance of ASV systems can be improved because the amount and variety of training data are increased. To combine NB and WB databases, a bandwidth extension (BWE) method is required. Recently, DNN-based BWE has been adopted as a data augmentation scheme for x-vector-based ASV systems [7]. In [7], the quality of WB signals generated from NB signals depended on the training data, and the DNN-based BWE method also required a large amount of training data.

Recently, the authors reported on the effectiveness of BWE methods for machine learning frameworks [11–13]. Additionally, the quality of speech generated by non-learning-based BWE is almost the same as that by learning-based BWE. Therefore, in this paper, non-learning-based BWE methods are employed for data augmentation for x-vector-based ASV systems. In the proposed scheme, non-learning-based BWE methods are adopted for NB databases. Extended and WB databases are used to estimate x-vector extractors. Since extended databases have different characteristics from those of WB, not only the amount of data but also the variety of data can be increased. To show the effectiveness of the proposed scheme, an experiment was conducted with using a variety of standard x-vector-based systems using VoxCeleb and speakers in the wild (SITW) databases. The results of using BWE methods showed that a proposed systems using the scheme reduced error 22.7% compared with our baseline system.

Section 2 of this paper introduces the state-of-the-art ASV system based on x-vector used in our experiment. Section 3 describes blind BWE methods and data augmentation, and section 4 illustrates our experimental setup and results. Finally, section 5 concludes the paper.

2. ASV system based on x-vector

2.1. X-vector and data augmentation

Recently, the x-vector-based ASV system [6] has been regarded as one of the state-of-the-art and de facto standard systems. Figure 1 shows a block diagram of the system. A DNN is constructed to map variable-length utterances into fixed-dimensional vectors. In the DNN architecture, the output vector from an embedding layer uses a speaker expression vector called an "x-vector." From [6, 7], x-vector-based ASV systems require a huge amount of training data for the highest performance. To prepare such a huge amount of training data, various data augmentation schemes have been proposed [8–10, 14].

2.2. Probabilistic linear discriminant analysis (PLDA)

PLDA-based back-end approaches have been proposed to reduce acoustic fluctuations and improve the performance of ASV systems [5]. In x-vector-based ASV systems, a PLDA model is estimated from speaker independent data. On PLDA-based frameworks, an vector ω_u extracted from an utterance u is assumed to be an observation from a probabilistic generative

10.21437/Odyssey.2020-64

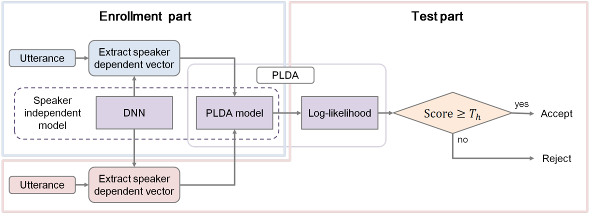

Figure 1: Block diagram of training x-vector-based ASV system and data augmentation with BWE

model as

$$\omega_u = \bar{\omega} + \Phi\delta + \Gamma\zeta_u + \epsilon_u, \quad (1)$$

where Φ and Γ are basis matrices that span speaker and channel subspaces. δ and ζ_uu express channel and speaker factors as standard Gaussian distributions. ϵ_u expresses residual error and follows a Gaussian distribution $N(\omega; 0, I)$, the mean vector of which is $0 \in R^{CD_F}$ and the covariance matrix $\Sigma \in R^{CD_F \times CD_F}$. $\bar{\omega}$ is offset in x-vector space. In equation (1), a probability generation model is defined as:

$$p(\omega_u|\delta, \zeta_u) = N(\bar{\omega} + \Phi\delta + \Gamma\zeta_u, \Sigma). \quad (2)$$

Below, from equation (2), is the log likelihood ratio for the hypothesis that calculates whether ω_1 and ω_2 are generated from the same speaker model (H_1) or not (H_0) using the x-vector ω_1 of the enrollment speaker and x-vector ω_2 of the evaluation speaker.

$$\log \frac{p(\omega_1, \omega_2|H_1)}{p(\omega_1|H_0)p(\omega_2|H_0)} \quad (3)$$

The log-likelihood ratio is used as an evaluation score.

3. Data augmentation with bandwidth extension

This paper aims to investigate the effectiveness of non-learning-based blind BWE methods for x-vector-based ASV systems.

3.1. Data augmentation

Since x-vector-based ASV systems use DNN architectures, they require a large amount of training data. In particular, a large amount of WB training data is required for the performance of x-vector-based ASV systems to be high. However, since there are few WB databases, the amount and the variety of data are limited. If BWE methods can be used to contribute a part of the training data for x-vector-based systems, the issue with data amount and variety would be relaxed. In some reports, the use of WB data extended from NB data by adding noise or by DNN-based BWE has already been considered [6, 7, 9]. However, the performance of learning-based BWE methods also depends on their training data, and it is difficult to reconstruct missing components due to bandwidth limitations. Recently, the authors reported on the effectiveness of non-learning-based BWE methods for machine learning frameworks. Additionally, the quality of speech generated by non-learning methods is almost the same as by learning-based BWE. In this paper, we aim to investigate the possibility of using non-learning-based BWE methods as a data augmentation scheme. In addition, we examine the effects of using a variety of training data on the system.

3.2. Linear prediction based analysis-synthesis (LPAS)

LPAS [15] is a blind and non-learning BWE method. The algorithm is based on a classical source filter model. A spectral envelope and residual error information are extracted from a NB signal by using linear prediction analysis. LPAS can relax the discontinuity of a power spectrogram, and the generated signals have high intelligibility.

3.3. Non-linear bandwidth extension (N-BWE)

N-BWE has been proposed as a blind and non-learning BWE [11–13]. In the N-BWE method, a non-linear function with two parameters and some filters is used to generate harmonics. This framework is very simple; despite that, it has been reported that the method has performed well with i-vector and x-vector-based ASV systems. From [12], signals generated by N-BWE obtained worse intelligibility scores than those of LPAS; however, the spectrograms of signals generated by N-BWE were closer to the target spectrogram than LPAS.

4. Experiments

4.1. Dataset description

X-vector-based ASV systems were by constructed according to the SITW v2 recipe in the Kaldi toolkit [16]. In the original recipe, a VoxCeleb database and SITW one were used for training and evaluation, respectively.

The VoxCeleb database was used to train DNNs as x-vector extractor and PLDA models [17, 18]. There were two datasets in the VoxCeleb database: VoxCeleb1 [17] and VoxCeleb2 [18]. The datasets were collected from interview videos uploaded to YouTube. VoxCeleb1 contained 153,516 utterances from 1,251 celebrities. VoxCeleb2 contained 1,092,009 utterances from 5,994 celebrities. These datasets were organized to include various ethnicities, occupations, ages and accents.

The SITW database [19] was used for the enrollment and test parts. The database was composed of utterances recorded without controlling situations such as the recording conditions or noise environments. The enrollment portion of SITW included 1,958 utterances from 119 speakers. The test one included 2,883 utterances for 180 speakers. Although SITW and VoxCeleb were collected separately, 60 speakers overlapped in the two databases. These speakers were deleted from both databases.

Two noise databases, MUSAN [20] and RIRNOISE [21], were used for data augmentation with noise. MUSAN consisted of music, noise, and speech signals. It contained over 900 noise signals, 42 hours of music from various genres, and 60 hours of speech from 12 languages. RIRNOISE consisted of three databases: point source-noises, real-rirs-isotropic-noises, and simulated-rirs. We used only simulated-rirs. The VoxCeleb, SITW, and the noise databases were sampled at 16 kHz.

For data augmentation with BWE, the NIST speaker recognition evaluation (SRE) 2005 [22] and NIST SRE 2006 [23] datasets were used. Through the BWE process, augmented data were generated since NIST SRE 2005 and 2006 were originally sampled at 8 kHz. The training datasets of NIST SRE 2005 and NIST SRE 2006 contained 1,492 and 10,468 utterances, respectively.

4.2. Conditions

For acoustic features, we used 30 order MFCCs including a log energy computed over a window of 25ms with a frame shift

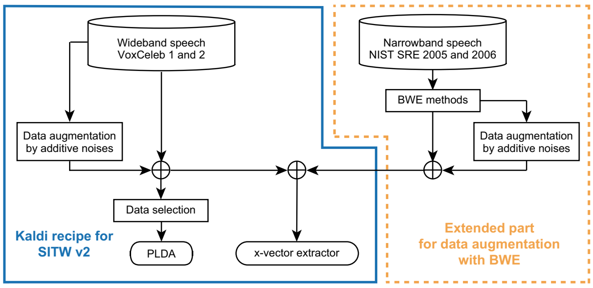

Figure 2: Block diagram of training x-vector-based ASV system and data augmentation with BWE

Table 1: Amount of training data for each system (utterances)

	x-vector extractor		PLDA
	VoxCeleb	NIST SRE	VoxCeleb
(A)	2,245,525	0	200,000
(B)	561,381	59,800	200,000
(C)	561,381	59,800	200,000
(D)	561,381	59,800	200,000
(E)	561,381	119,600	200,000
(F)	561,381	119,600	200,000
(G)	561,381	119,600	200,000
(H)	561,381	119,600	200,000
(I)	561,381	179,400	200,000

of 20ms. Delta and acceleration were appended to each static feature vector. In Fig. 2, the part surrounded by the blue line shows a block diagram of the original recipe for training x-vector-based ASV systems with the Kaldi toolkit. In the original recipe, WB data, which included 1,245,525 utterances from the VoxCeleb database, was prepared for training an x-vector extractor. Data augmented by adding noise was generated from the VoxCeleb data. Through the augmentation with noise, over 4,000,000 utterances were generated, and 1,000,000 utterances were randomly selected. The total amount of training data for the x-vector extractor was 2,245,525 utterances. For training a PLDA model, 200,000 utterances of the longest duration from all of the training data for the x-vector extractor were selected.

The part surrounded by the yellow dotted line in Fig. 2 illustrates a block diagram of data augmentation with BWE methods. This proposed part was regarded as an extension of the original recipe, and NIST SRE 2005 and NIST SRE 2006 were used for NB data. In the extended part, data augmentation with noise was also performed for the data augmented by BWE. Thus, the issue with data amount and variety for x-vector-based ASV systems was expected to be relaxed. The following systems were compared.

(A) 16k

This system was constructed under the original x-vector recipe of the Kaldi toolkit with all VoxCeleb and SITW data. This was regarded as a case of using a large amount of WB data.

(B) 16k(quarter)

This system was constructed in the same manner as the original x-vector recipe. In this condition, the amount of WB data was reduced from 1,245,525 to 311,381 utterances, and the amount of data augmented by adding noise was 250,000 utterances. The total amount of training data was 561,381 utterances. In this paper, this system was regarded as the baseline system.

(C) DA(UP)

Without training the x-vector extractor, the manner of the system construction was the same as (B). To train the x-vector extractor, data augmented by simple upsampling were added to the training data of the baseline system. The total amount of training data was 621,181 utterances. The simple upsampling was performed with an upsampling factor 2 and a low-pass filter.

(D) DA(N-BWE)

The procedure for constructing this system was the same as (C). Instead of simple upsampling, data augmented by N-BWE [11] was added to the training data of the baseline system. The total amount of training data was 621,181 utterances. The parameter settings of the N-BWE were the same as [11].

(E) DA(LPAS)

The procedure for constructing this system was the same as (C). Instead of simple upsampling, the data augmented by LPAS [15] was added to the training data of the baseline system. The total amount of training data was 621,181 utterances. The filters settings of LPAS were the same as [15].

(F) DA(UP&N-BWE)

The procedure for constructing this system was the same as (C). In the extended part, simple upsampling (C) and N-BWE (D) were used for data augmentation with BWE. The total amount of training data was 742,362 utterances.

(G) DA(UP&LPAS)

The procedure for constructing this system was the same as (C). In the extended part, simple upsampling (C) and LPAS (E) were used for data augmentation with BWE. The total amount of training data was 742,362 utterances.

(H) DA(N-BWE&LPAS)

The procedure for constructing this system was the same as (C). In the extended part, N-BWE (D) and LPAS (E) were used for data augmentation with BWE. The total amount of training data was 742,362 utterances.

(I) DA(UP&N-BWE&LPAS)

The procedure for constructing this system was the same as (C). In the extended part, simple upsampling (C), N-BWE (D), and LPAS (E) were used for data augmentation with BWE. The total amount of training data was 802,162 utterances.

For the above systems, the amount of data augmented by adding noise was quadrupled from that of the original NB data. Table 1 shows the amount of training data for each system. Even though the amounts of training data (A) and (B) for PLDA were the same, the included utterances were different. All methods using data augmentation with BWE (C)-(I) utilized the same PLDA model as (B).

All experiments were evaluated with an equal error rate (EER) and a minimum detection cost function (minDCF) [24]. EER assigns equal weights to a false negatives rate (FAR) and a false positives rate (FRR). MinDCF is usually used to assess performance in settings where achieving a low FRR is more important than achieving a low FAR. The difference between minDCF IE-2 and minDCF IE-3 is whether the P-targets of parameter is 0.01 or 0.001. These parameters

Table 2: EER (%) and minDCF for each system

x-vector systems	SITW Core task		
conditions	EER	minDCF IE-2	minDCF IE-3
(A) 16k	3.554	0.3636	0.5296
(B) 16k(quarter)	6.616	0.5722	0.7862
(C) DA(UP)	5.221	0.4943	0.7139
(D) DA(N-BWE)	5.358	0.5031	0.7249
(E) DA(LPAS)	**5.112**	0.4932	**0.6838**
(F) DA(UP&N-BWE)	5.139	0.4817	0.6961
(G) DA(UP&LPAS)	**5.112**	**0.4556**	**0.6913**
(H) DA(N-BWE&LPAS)	5.221	**0.4787**	0.6979
(I) DA(UP&N-BWE&LPAS)	5.522	0.5287	0.7564

were defined in the NIST SRE evaluation plans.

4.3. Results

Table 2 shows the EER and minDCF of the x-vector-based systems for each system. Comparing (A) 16k with (B) 16k (quarter), the EER and minDCF scores of (B) increased significantly since the amount of training data was reduced to a quarter of (A). This indicates that the amount of training data affected the performance of the x-vector-based ASV systems seriously. Comparing (B) 16k (quarter) with the systems using data augmentation with BWE (C)-(I), the EERs of all systems using data augmentation with BWE were lower than that of (B). It is thought that the performance of the x-vector-based ASV systems was improved by adopting the data augmentation with BWE. Comparing the systems using single BWE methods, (C) DA(UP), (D) DA(N-BWE), and (E) DA(LPAS), while the same amount of data was used, the performances were different. System (E) achieved the lowest EER and reduced error 22.7% compared with our baseline system. Comparing (F) DA(UP&N-BWE), (G) DA(UP&LPAS), and (H) DA(N-BWE&LPAS), although these systems had the same amount of training data, (F) and (G) had the highest performance. One of the reasons was that the harmonics of the WB speech generated with LPAS and N-BWE contained a similar variety of data. The variety was different from that of simple upsampling. Therefore, the combination of UP and N-BWE (F) or UP and LPAS (G) was almost the same in terms of data variation, and the performance was almost the same. The combination of LPAS and N-BWE did not result in high performance since the characteristics of LPAS and N-BWE were similar and the variety of data was not expanded. (I) DA(UP&N-BWE&LPAS) had a lower EER than (B); however, when comparing (I) with the other data augmentation systems, the improvement was the lowest. The results show that the effect of data augmentation on x-vector-based systems depends not only on the amount of data but also the variety of the data. Furthermore, comparing (E) with (G), which had the lowest EER, (E) yielded the lowest minDCF IE-3, but the score of minDCF IE-2 was slightly improved. On the other hand, the minDCF IE-2 of (G) was the lowest, and the minDCF IE-3 of (G) was the second lowest among the systems with data augmentation with BWE. This indicates that the performance of (G) was more stable than that of (F). From the results, the performance was improved by considering data variety and increasing the data amount.

We also applied data augmentation to PLDA training, but this did not improve the performance. Additionally, we performed preliminary experiments to apply data augmentation for (A), and the EER was improved.

5. Conclusion

In this paper, we proposed a data augmentation scheme with BWE for DNN-based ASV systems. One approach used with these systems is named, x-vector, and it requires a large amount of training data. However, since there are few WB databases, the amount and the variety of data are limited. If BWE methods can be used to provide a part of the training data for x-vector-based systems, the issue with data amount and variety is relaxed. In this paper, we aimed to demonstrate that several non-learning-based BWE methods are useful for data augmentation for x-vector-based ASV systems. By using SITW and NIST SRE databases, experimental results showed that the a system using the proposed scheme reduced error 22.7% compared with our baseline system.

Future work will involve changing the variation of data by applying other NB databases published by NIST SRE and used for x-vector-based ASV systems. Additionally, a different training method will be considered.

6. Acknowledgment

This work was supported, in part, by JSPS KAKENHI Early-Career Scientists Grant number JP19K20271 and ROIS-DS-JOINT (021RP2019) to S. Shiota.

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *in Proc. INTERSPEECH*, pp. 999–1003, 2017.

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

[4] W. Hsu, Y. Zhang, R. J. Weiss, Y. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," *in Proc. ICASSP*, pp. 5901–5905, 2019.

[5] S. J. D Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *in Proc. ICASSP*, pp. 1–8, 2007.

[6] D. Snyder, D. Garcia-Romero, G. Shell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *in Proc. ICASSP*, 2018.

[7] P. S. Nidadavolu, V. Iglesias, J. Villalba, and N. Dehak, "Investigation on neural bandwidth extension of telephone speech for improved speaker recognition," *in Proc. ICASSP*, pp. 6111–6115, 2019.

[8] C. Chen, S. Zhang, C. Yeh, J. Wang, T. Wang, and C. Huang, "Speaker characterization using tdnn-lstm based speaker embedding," *in Proc. ICASSP*, pp. 6211–6215, 2019.

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

[10] Z. Wu, S. Wang, Y. Qian, and K. Yu, "Data augmentation using variational autoencoder for embedding based speaker verification," *in Proc. INTERSPEECH*, pp. 1163–1167, 2019.

[11] H. Miyamoto, S. Shiota, and H. Kiya, "Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts," *in Proc. APSIPA Annual Summit and Conference*, pp. 1868–1874, 2018.

[12] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Blind bandwidth extension with a non-linear function and its evaluation on automatic speaker verification," *IEICE trans. Inf. Sys*, vol. E103-D, pp. 42–49, 2019.

[13] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation on x-vector-based speaker verification," *in Proc. INTERSPEECH*, pp. 4055–4059, 2019.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," *in Proc. ICASSP*, pp. 5796–5800, 2019.

[15] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," *in Proc. ICASSP*, pp. 5429–5433, 2018.

[16] P. Daniel, G. Arnab, B. Gilles, B. Lukas, G. Ondrej, G. Nagendra, H. Mirko, M. Petr, Q. Yanmin, S. Petr, et al., "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[17] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[19] M. Mitchell, F. Luciana, C. Diego, and L. Aaron, "The Speakers in the Wild (SITW) speaker recognition database," *in Proc. INTERSPEECH*, pp. 818–822, 2016.

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

[21] 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 Proc. ICASSP*, pp. 5220–5224, 2017.

[22] "Nist (2005) the nist year 2005 speaker recognition evaluation plan," `https://catalog.ldc.upenn.edu/docs/LDC2011S01/sre-05_evalplan-v5.pdf`, 2004.

[23] "The nist year 2006 speaker recognition evaluation plan," `https://catalog.ldc.upenn.edu/docs/LDC2011S09/sre-06_evalplan-v9.pdf`, 2006.

[24] "Nist 2016 speaker recognition evaluation plan," `https://www.nist.gov/system/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf`, 2016.