IEICE TRANS. INF. & SYST., VOL.E103–D, NO.1 JANUARY 2020

PAPER Special Section on Enriched Multimedia—Application of Multimedia Technology and Its Security—

Blind Bandwidth Extension with a Non-Linear Function and Its Evaluation on Automatic Speaker Verification

Ryota KAMINISHI[†a)], **Haruna MIYAMOTO**[†b)], ***Nonmembers***, **Sayaka SHIOTA**[†c)], ***Member***, ***and*** **Hitoshi KIYA**[†d)], ***Fellow***

SUMMARY This study evaluates the effects of some non-learning blind bandwidth extension (BWE) methods on state-of-the-art automatic speaker verification (ASV) systems. Recently, a non-linear bandwidth extension (N-BWE) method has been proposed as a blind, non-learning, and light-weight BWE approach. Other non-learning BWEs have also been developed in recent years. For ASV evaluations, most data available to train ASV systems is narrowband (NB) telephone speech. Meanwhile, wideband (WB) data have been used to train the state-of-the-art ASV systems, such as i-vector, d-vector, and x-vector. This can cause sampling rate mismatches when all datasets are used. In this paper, we investigate the influence of sampling rate mismatches in the x-vector-based ASV systems and how non-learning BWE methods perform against them. The results showed that the N-BWE method improved the equal error rate (EER) on ASV systems based on the x-vector when the mismatches were present. We researched the relationship between objective measurements and EERs. Consequently, the N-BWE method produced the lowest EERs on both ASV systems and obtained the lower RMS-LSD value and the higher STOI score.

key words: *blind bandwidth extension, non-linear function, automatic speaker verification, i-vector, x-vector*

1. Introduction

Automatic speaker verification (ASV) refers to a technique that uses voices to identify people. Recent state-of-the-art ASV techniques include i-vector approach [1], [2], probabilistic linear discriminant analysis (PLDA) classifier [3], and methods based on the x-vector [4]–[6]. Thanks to these methods, the performance of ASV systems has dramatically improved with narrowband (NB) or wideband (WB) databases, such as the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) [7] or Speaker In The Wild (SITW) [8]. The state-of-the-art ASV systems require a large amount of training data for obtaining high performance, and data augmentation is regarded as an important factor for ASV performance. It is well known that almost databases released by NIST SRE series are sampled at 8 kHz (NB) and the SITW database is sampled at 16 kHz (WB). Additionally, some databases sampled at 32, 48 kHz, and so on. Therefore, sampling mismatch problem has already happened and discussed in ASV research area [9]–[11]. Moreover, some applications using voice patterns adopt client server system (CSS). These CSSs are required to assume many communications technology which including several bandwidth limitations and recording environments. Depended on these assumptions, sampling mismatch problems are caused in some steps of these systems. When mismatches are present, data that has a higher sampling rate is usually downsampled to a lower one [12]. However, downsampling all training data and reconstructing the ASV systems is expensive. It is well-known that a lower sampling rate causes the ASV performance to decline [9]–[11]. Bandwidth extension (BWE) methods can be used to correspond lower sampling rates to higher ones.

BWE methods are regarded as methods for restoring high-frequency losses caused by band limits [9], [10], [13]–[17]. Many BWE approaches have already been reported, and they are categorized into blind or non-blind methods. Non-blind methods restore missing frequency components from auxiliary high-frequency (HF) side information encoded into a data stream together with low-frequency (LF) components. In contrast, blind methods use only the LF components to estimate missing HF components. A recently non-linear BWE (N-BWE) method took a blind, non-learning, and light-weight BWE approach [9]. It performed well in terms of speaker individuality and root mean square log-spectral distortion (RMS-LSD). Additionally, non-learning BWE approaches are also reported [9], [17]–[20] in recent years.

Although it has been reported that some ASV approaches estimate models with NB and WB mixed data [10], few studies have investigated the effects of applying non-learning BWE methods to ASV systems. For training i-/x-vector-based ASV systems, there are three portions of dataset: for training speaker independent (SI) models, for estimating enrollment vectors and for evaluation. Therefore, we assume two scenarios as sampling mismatch problems. One is that the data for SI models are sampled at WB conditions, but the enrollment and the evaluation data are sampled at NB conditions. The other one is that the data for SI and enrollment models are sampled at WB conditions, but the evaluation data is sampled at NB conditions. These mismatch problems are depended on systems and this problem will also face between WB and super WB conditions. Since the non-learning BWE methods have some possibilities to relax the mismatch problems, this paper investigates their

Manuscript received March 29, 2019.
Manuscript revised August 9, 2019.
Manuscript publicized October 25, 2019.

[†]The authors are with Tokyo Metropolitan University, Hinoshi, 191–0065 Japan.
a) E-mail: kaminishiryota4869@gmail.com
b) E-mail: miyamoto-haruna@ed.tmu.ac.jp
c) E-mail: sayaka@tmu.ac.jp
d) E-mail: kiya@tmu.ac.jp

DOI: 10.1587/transinf.2019MUP0008

effectiveness.

This paper is focused on the non-learning BWE methods and the effects they have on i-/x-vector-based ASV systems. To evaluate the effectiveness of the BWE methods, we carried out an i-/x-vector-based ASV experiment and some objective evaluations. Consequently, the N-BWE method produced the lowest equal error rate (EER) and obtained one of the lowest RMS-LSD values and the higher STOI scores from the SITW dataset.

Section 2 of this paper introduces the state-of-the-art ASV systems under in our experiment. Section 3 describes non-linear bandwidth extension, and Sect. 4 illustrates our experimental setup and the results. Finally, Sect. 5 concludes the paper.

2. Automatic Speaker Verification Systems

In this section, two ASV systems based on i-vector and x-vector are described as state-of-the-art systems. The block diagram of a basic ASV system with a Gaussian-PLDA approach is depicted in Fig. 1. Based on this diagram, both ASV systems were built using the Kaldi toolkit [21].

2.1 I-Vector

As one of the state-of-the-art ASV system, i-vector-based ASV system has been reported [1]. By using factor analysis, a low-dimensional vector containing speaker individuality is extracted from a supervector mean M_u for given utterance u as follows:

$$M_u = m_{ubm} + T\omega_u, \tag{1}$$

where $m_{ubm} \in R^{CD_F}$ and $T \in R^{CD_F \times D_T}$ are called a Gaussian Mixture Model (GMM) supervector of a universal background model (UBM) and a total variability (TV) matrix, respectively. C is the number of mixture components, and D_F is the dimension of acoustic features. $\omega_u \in R^{D_T}$ is a latent variable for the utterance u, and it is called "i-vector." D_T represents the dimension of i-vector. ω_u follows a Gaussian distribution $N(\omega; 0, I)$ whose mean vector is $0 \in R^{D_T}$ and the covariance matrix is an identify matrix $1 \in R^{D_T \times D_T}$.

2.2 X-Vector

A recent ASV system based on the x-vector is another recently developed state-of-the-art system called "x-vector" [22]. Speaker individuality is represented by DNN embeddings [23]. The DNN structure is shown in Fig. 2. The inclusion of i_s^t means that feature vectors are extracted from an utterance s and frame $t = \{1, \cdots, T\}$. The x-vector that represents the speaker is extracted from the embedding layers. The second or third layers of the DNN structure in Fig. 2 works with the framewise input features. The statistics pooling layer aggregates all T frame-level outputs from previous layer and computes its mean and standard deviation. The embedding layers (Emb) are trained with segment-wise features through the pooling layer.

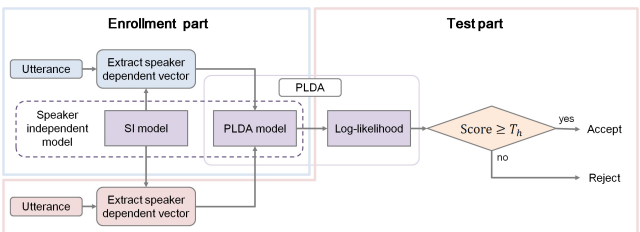


Fig. 1 Block diagram of ASV system

2.3 Gaussian PLDA

For ASV back-end systems, a Gaussian-PLDA (G-PLDA) classifier is used [24]. On G-PLDA-based frameworks, an extracted vector ω_u from an utterance u, is assumed to be an observation from a probabilistic generative model as

$$\omega_u = \bar{\omega} + \Phi\delta + \Gamma\zeta_u + \epsilon_u, \tag{2}$$

where Φ and Γ are basis matrices that span speaker and channel subspace. δ and ζ_u 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 $0 \in R^{D_T}$ and the covariance matrix $\Sigma \in R^{CD_F \times CD_F}$. $\bar{\omega}$ is a global offset. In Eq. (2), the probability generation model is defined as follows,

$$p(\omega_u|\delta, \zeta_u) = N(\bar{\omega} + \Phi\delta + \Gamma\zeta_u, \Sigma). \tag{3}$$

When the vectors of enroll speaker ω_1 and test speaker ω_2 are obtained, an identification score s(ω_1, ω_2) is calculated as the log-likelihood ratio for hypothesis test, which was in the same speaker model (H_1) and in the different speaker models (H_0) as shown below,

$$s(\omega_1, \omega_2) = \log \frac{p(\omega_1, \omega_2|H_1)}{p(\omega_1|H_0)p(\omega_2|H_0)}. \tag{4}$$

The G-PLDA-based back-end approach can reduce the acoustic fluctuation and improve ASV system performance.

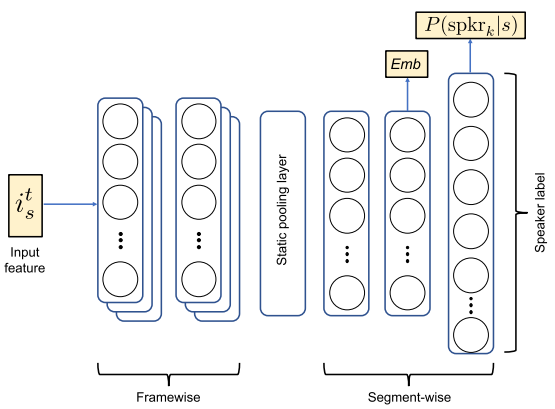


Fig. 2 DNN structure for x-vector

3. Bandwidth Extension Methods

3.1 Categories of BWE Methods

In the last decade, many bandwidth extension (BWE) methods have been developed [9], [10], [18], [25]. These approaches can be categorized into blind or non-blind and non-learning or learning. Non-blind approaches must reserve some bandwidth for additional information, which helps to restore missing information by controlling the bandwidth. However, received servers have to change their decoding protocols for non-blind BWE methods. Blind approaches restore the missing information without providing any additional information. Almost all research focuses on the blind approach because it requires no change to the decoding protocols. For the other category, many learning approaches are reported [10], [25]–[27] thanks to the development of machine learning techniques. Learning approaches require a large amount of speech data and hard parameter tuning to train accurate models. Non-learning approaches have also been developed [9], [17]–[19]. Non-learning methods focus on situations that involve lightweight processing and constraint-free amounts of data. In this paper, we focus on a blind and "non-learning" BWE approach.

3.2 Spectrum Shifting

The spectrum shifting method (SHIFT) was reported in [28]. After a basic upsampling with a low-pass filter and an interpolator factor, this method modulates the period under $F_{s_0}/2$ [Hz] for generating high frequency components. A WB signal can be obtained by filling the free frequency domain $(F_{s_0}/2 - F_{s_1}/2)$ [Hz]. Here, F_{s_0} and F_{s_1} are original sampling rate and upsampled rate, respectively. When m is an upsampling factor, $F_{s_1} = mF_{s_0}$.

3.3 Linear Prediction-Based Analysis-Synthesis

Linear prediction-based analysis-synthesis (LPAS) was developed in [18] as a SHIFT-based method. This algorithm is based on a classical source-filter model. Spectral envelope and residual error information is extracted from an NB signal by using linear prediction analysis. The generated high-frequency components are more natural than the ones generated by SHIFT.

3.4 N-BWE

An N-BWE method has been proposed as a blind and non-learning BWE approach [9]. Figure 3 shows the block diagram of the N-BWE method. By using basic upsampling, an upsampled signal $y_{UP}[n]$ is generated. n is a discrete-time variable. $y_{UP}[n]$ has no harmonic components. Before passing the non-linear function, the upsampled signal $y_{up}[n]$ is convoluted a filter $h_A[n]$ to select the bandwidth to generate

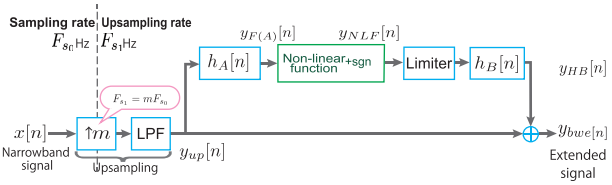


Fig. 3 Block diagram of non-linear BWE method

harmonics. The signal denotes $y_{F(A)}[n]$. A non-linear function can be used to generate harmonic components, and a general form is given by

$$y_{NLF}[n] = \mathrm{sgn}(y_{F(A)}[n]) \cdot |y_{F(A)}[n]^{\alpha}| \times \beta, \tag{5}$$

with

$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}, \tag{6}$$

where α and β are the parameters for controlling the nonlinearity, and a is a real value. Here, α and β control the frequency of the generated harmonics and the magnitude of the generated components, respectively. Since these parameters affect the quality of extended signal $y_{bwe}[n]$ non-linearly, the speaker verification performance sometimes depends on the parameters. There are some approaches to decide the parameter values by trying many parameters or solving some optimization problems. This paper focuses on investigating the abilities of N-BWE, the parameters uses the same of [9] which have already searched the adequate parameters. To control the bandwidth of $y_{F(A)}[n]$, in this paper, the impulse response of a digital filter, $h_A[n]$ in Fig. 3, is assumed to be an all pass filter.

$$h_A[n] = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}. \tag{7}$$

The limiter in Fig. 3 is given as,

$$y_{HB}[n] = \begin{cases} y_{NLF}[n], & y_{NLF}[n] \leq T_h \\ M, & y_{NLF}[n] > T_h \end{cases}, \tag{8}$$

where T_h is a threshold value and M is a constant value. The parameters T_h and M control clipping conditions of input signals. Basically, these parameters set to the same value, and it depends on encoding methods or using software. Finally, in order to reduce aliasing artifacts, the digital filter $h_B[n]$ is used. For the digital filters $h_A[n]$ and $h_B[n]$, a high-pass, a low-pass or a band-pass filter can be used. On the basis of procedure in Fig. 3, it is expected that $y_{HB}[n]$, will compensate for high-frequency losses. Based on the procedure in Fig. 3, it is expected that $y_{HB}[n]$ will compensate for high-frequency losses.

3.5 Spectrogram Comparison of Each Method

Figure 4 shows spectrogram examples of speech signals.

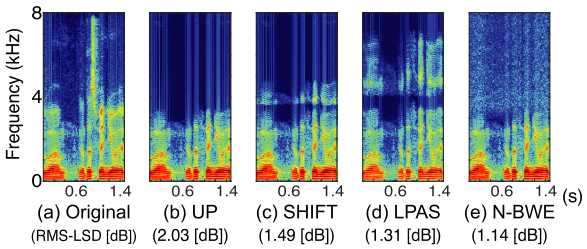


Fig. 4 Spectrogram examples of speech signals ($m = 2$; $F_{S_0} = 8$kHz, $F_{S_1} = 16$kHz)

First, the original signal (a) sampled at 16 kHz has frequency components from 0 kHz to 8 kHz. The upsampled signal (b) from 8 kHz to 16 kHz contains only low frequency components under 4 kHz. Signal (c) was generated by SHIFT, signal (d) was generated by LPAS, and signal (e) was generated by N-BWE. As these examples show, the BWE methods in Figs. 4 (c), (d) and (e) can generate harmonic components in high-frequency components. The root mean square log spectral distortion (RMS-LSD) scores are also shown in Fig. 4. The lower the RMS-LSD score, the closer the degraded speech sample is to its reference. Even though the spectrogram of N-BWE showed a low similarity, N-BWE can generate the harmonics with smoothing effects. Consequently, it can be seen that the RMS-LSD score of N-BWE was the lowest of all BWE methods.

4. Experiments

To evaluate the effectiveness of the non-learning BWE methods, we carried out ASV experiments based on i-vector and x-vector and some objective evaluations.

4.1 Database

The Kaldi toolkit [21] and a recipe for the Speaker In The Wild (SITW) database [8] were used to construct an ASV system. The Voxceleb dataset was used to estimate the DNN and G-PLDA. There were two versions of this dataset: Voxceleb 1 [29] and Voxceleb 2 [30]. The databases were collected from interview videos uploaded to YouTube. Voxceleb 1 contained over 100,000 utterances from 1,251 celebrities. Voxceleb 2 contained over 1,000,000 utterances from 6,112 celebrities. In both versions, the speakers spanned a wide range of different ethnicities, accents, professions, and ages. Their nationalities and genders were provided as well. The evaluation task was performed using the SITW database, which contained 299 speakers. The SITW database was split into two tasks. It named development task and evaluation one. There was no speaker belongs to both tasks. The development task contained 2,597 target and 335,629 impostor trials from 119 unique speakers and the evaluation task contained 3,658 target and 718,130 impostor trials from 180 unique speakers. Unlike existing databases for ASV systems, this data was not recorded under controlled conditions and contained real noise. We tested each method on the core-core task of the Kaldi recipe for SITW. Although SITW and Voxceleb were collected independently, there was an overlap of 60 speakers in both datasets. The overlapping speakers were removed from Voxceleb prior to using them as training data. Two noise databases were used for data augmentation. One was MUSAN [31], which consisted of music, noise, and speech. It contained over 900 noise signals, 42 hours of music from various genres and 60 hours of speech from 12 languages. We used this database for adaptive noise to Voxceleb. Another one was RIRNOISE [32], which consisted of three database: pointsource-noises, real-rirs-isotropic-noises, and simulated-rirs. We used only simulated-rirs. All databases were sampled at 16 kHz.

4.2 Experimental Conditions

We assumed two scenarios. The first was that the sampling rate mismatch was caused by the enrollment and the test data (Test). The second was that the training data of the speaker independent models was of a higher sampling rate, and the enrollment and test data were of a lower sampling rate (Enroll). In both scenarios, the WB signals were sampled at 16 kHz and the NB signals were sampled at 8 kHz. Equal error rate (EER) was used as an evaluation measurement. EER is the point where the false rejection rate (FRR) and the false acceptance rate (FAR) become equal, the lower the value, the better the accuracy. For objective evaluation, perceptual evaluation of speech quality (PESQ), short-time objective intelligibility measure (STOI), and RMS-LSD were used. PESQ and STOI represented the naturalness of degraded speech by comparing with a reference one. The PESQ score ranged from 0 (bad) to 4.5 (best). The STOI value ranged from 0.0 (bad) to 1.0 (best). RMS-LSD measured the log spectral distance between a degraded piece of speech and a reference one.

For i-vector-based systems, standard mel-frequency cepstrum coefficients (MFCCs) extraction was used as acoustic feature. In the feature extraction, we used 24-order MFCCs computed over a window of 25 ms with a frame shift of 20 ms. The UBM had 2048 Gaussian mixtures components. The system used a 400 dimensional i-vector extractor and G-PLDA scoring. Although SI models, such as UBM, TV matrix and G-PLDA, were trained with Voxceleb 1 and Voxceleb 2, training models required too much time because both databases contained over 1,000,000 utterances. Therefore, UBM and TV matrix were trained after reducing them from 1,000,000 to 100,000 utterances as original data. For x-vector-based systems, we used 30-order MFCCs with the same manner of the i-vector-based system. The DNNs and the G-PLDA models were trained 100,000 utterances in Voxceleb1 and Voxceleb 2 as original data and augmented data by using MUSAN and RIRs noise database. Table 1 shows the DNN architecture used by the x-vector systems. These conditions for the x-vector-based systems were set to the same as the original paper of x-vector [22].

Table 1 DNN architecture

Layer	Layer context	Total context	input x output
Frame1	{t-2, t+2}	5	120x512
Frame2	{t-2, t, t+2}	9	1536x512
Frame3	{t-3, t, t+3}	15	1536x512
Frame4	{t}	15	512x512
Frame5	{t}	15	512x1500
Stats pooling	[0,T)	T	1500Tx3000
Segment6	{0}	T	3000x512
Segment7	{0}	T	512x512
Softmax	{0}	T	512xS

Table 2 Experimental conditions for each method

Scenario	Condition	SI model	Enroll	Test
Mismatch scenario (Enroll)	(A) UP	Original (16 kHz)	UP	UP
	(B) SHIFT		SHIFT	SHIFT
	(C) LPAS		LPAS	LPAS
	(D) N-BWE		N-BWE	N-BWE
Mismatch scenario (Test)	(E) UP		Original (16 kHz)	UP
	(F) SHIFT			SHIFT
	(G) LPAS			LPAS
	(H) N-BWE			N-BWE
Matched cond.	(I) Down		8 kHz	8 kHz
	(J) Org		16 kHz	16 kHz

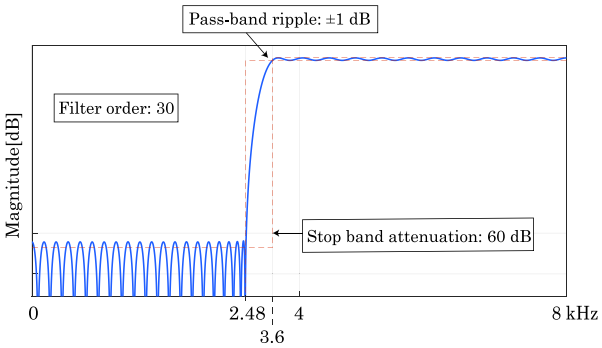


Fig. 5 Filters designed for N-BWE

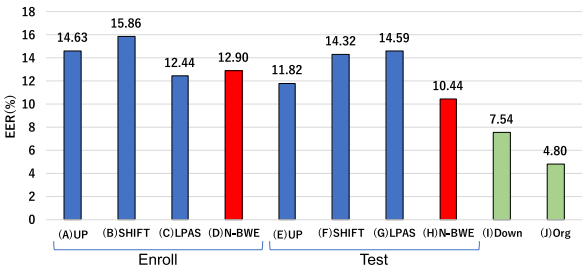


Fig. 6 EERs for each conditions on ASV systems based on i-vector (development task)

4.3 Comparison Conditions

Based on the two mismatched scenarios, comparison conditions were set as shown in Table 2. The details were denoted as follows.

(A) UP (Enroll)
The enroll and test data was simply upsampled. Note that the speech samples did not include any harmonic components in the high-frequency components.

(B) SHIFT (Enroll)
The enroll and test data was extended by SHIFT [17]. The band-pass filter was the same as [28].

(C) LPAS (Enroll)
All data for enroll and test was extended by LPAS [18] from the NB speech sampled at 8kHz.

(D) N-BWE (Enroll)
The enroll and test data was extended by using the N-BWE method [9] from the NB speech sampled at 8kHz. The optional filter $h_A[n]$ was defined as the all pass filter, and the filter $h_B[n]$ was defined as Fig. 5. To control the nonlinearity, α and β in Eq. (5) were set to 2 and 100,000, respectively.

(E) UP (Test)
For the enrollment data, original speech was used. The test data was upsampled from 8kHz to 16kHz.

(F) SHIFT (Test)
From (E), the test data was extended by SHIFT. The band-pass filter used [28].

(G) LPAS (Test)
From (E), the test data was extended by LPAS.

(H) N-BWE (Test)
From (E), the test data was extended by N-BWE. The filters and parameters were the same as (D).

(I) Down
All data was downsampled from 16 kHz to 8 kHz. This is denoted as NB signal $x[n]$ in Fig. 3.

(J) Org
All data was used without any modifications.

The parameters for N-BWE (α, β, T_h, M) were decided from the preliminary experiments with different databases. Thus, the parameters were not optimized for these experiments of this paper, but the parameters commonly fitted to many databases. Since it is expensive to reconstruct the UBM, TM matrix and G-PLDA models for each condition, the original data sampled at 16 kHz was used for the SI models. In the case of (I) Down, all data were downsampled at 8 kHz.

4.4 Results

4.4.1 I-Vector

Figures 6 and 7 show the EERs on the i-vector systems for each of the conditions under a development task and an evaluation one, respectively. From Fig. 6, comparing the EER of (I) with that of (J), when the sampling rate mismatch was not present, the ASV performance did not change significantly. However, when the mismatch was present, the EERs of (A) - (H) were considerably higher than those of (I) and (J). This result suggests that the sampling rate mismatches are still big problems. The results of the Enroll scenario and the Test scenario had a similar tendency. The EERs of (A) and (E) were high due to the missing information, al-

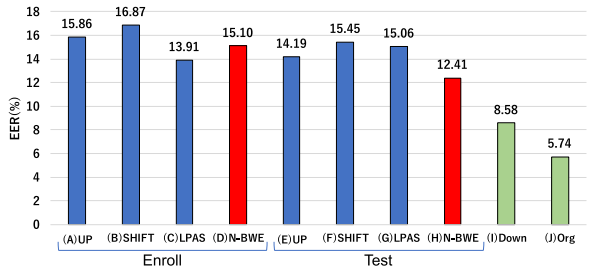


Fig. 7 EERs for each conditions on ASV systems based on i-vector (evaluation task)

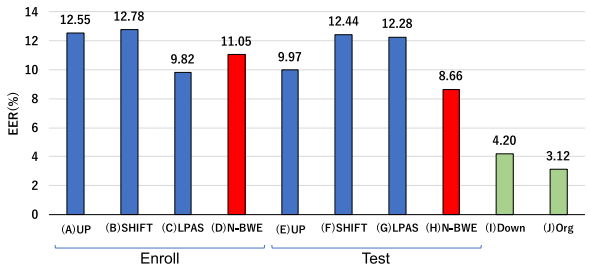


Fig. 8 EERs for each conditions on ASV systems based on x-vector (development task)

though the EERs of SHIFT-based methods (B) and (F) were obtained higher values. This means that SHIFT can generate WB components. However, the speaker individualities were not suitable. The LPAS and N-BWE conditions obtained lower EERs than (A) and (E). This proves that the non-learning BWE had some potential for reducing mismatch problems without the learning process. Both EERs of N-BWE achieved significantly lower EERs in both scenarios. The EERs of "test" scenario were lower than those of "Enroll" scenario. It can be considered that the sampling mismatch between the SI models and the enrollment utterances makes the accuracy of the enrollment models low. In the test scenario, the enrollment model can estimate adequately, and the BWE methods help to compensate for the sampling mismatch between the enrollment models and test utterances. From Fig. 7, it can be seen that almost all results were the same as Fig. 6. Thus, the BWE methods had the same effects for the different tasks.

4.4.2 X-Vector

Figures 8 and 9 show the EERs on the x-vector systems for each of the conditions under the development and evaluation tasks, respectively. Comparing the results of the i-vector with that of x-vector, the performance of x-vector worked well than that of i-vector. It showed the x-vector-based ASV systems work with the high performance. However, the sampling mismatches still caused to obtain considerably high EERs. The EER of all conditions had almost the same tendency as the i-vector results. Therefore, the potential of the BWE methods to reduce mismatch problems was not dependent on the ASV systems.

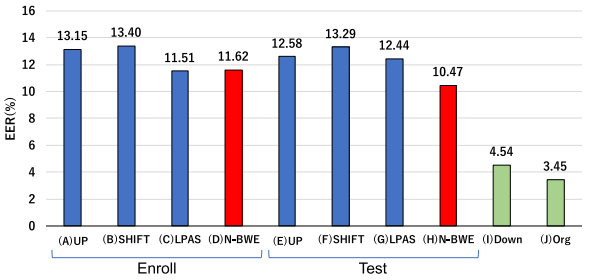


Fig. 9 EERs for each conditions on ASV systems based on x-vector (evaluation task)

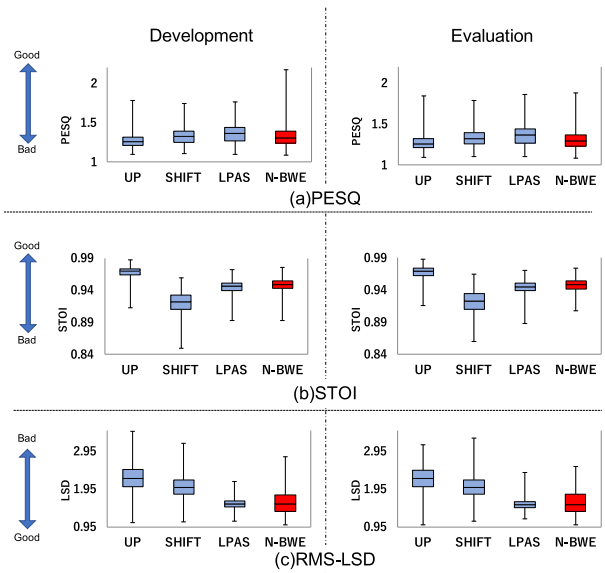


Fig. 10 Objective results for each BWE methods

4.4.3 Objective Results

Figure 10 illustrates the PESQ, STOI or RMS-LSD scores for each BWE method. Objective experiments were performed with all utterances of the SITW database. From the results, BWE methods have no precise advantage in terms of the objective measurements. However, N-BWE obtained slightly better scores than SHIFT and LPAS. Comparing the RMS-LSD scores with EERs of Figs. 6–9, the method which obtained the lower RMS-LSD tends to obtain the lower EER.

Consequently, N-BWE can help to compensate for the bandwidth limitation for state-of-the-art ASV systems.

5. Conclusion

This paper evaluated the effects of some non-learning and blind BWE methods on ASV systems based on i-vector and x-vector. The N-BWE is a blind, non-learning and lightweight BWE approach. Other non-learning BWE methods have also been developed in recent years. We investigated the influence of sampling rate mismatches and

the performance of BWE methods against mismatches. The N-BWE method improved the EER of ASV systems based on i-vector and x-vector. We researched the relationship between objective measurements and EERs. Consequently, the N-BWE method produced the lowest EER and obtained the lower RMS-LSD value and the higher STOI score.

In the future, the BWE methods will be evaluated with regards to the algorithmic delay. Because BWE methods generate amplitude information only, phase estimation will be adopted to make reconstructed signals more natural. We will also discuss about scoring approaches to evaluate the BWE methods. Additionally, since the BWE methods can use as a technique for data augmentation, the effectiveness will be evaluated.

Acknowledgments

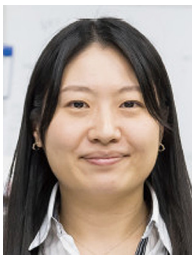
This work was supported in part by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) JP16757733, JSPS KAKENHI Early-Career Scientists Grant number JP19K20271, and ROIS-DS-JOINT (021RP2019) to S. Shiota.

References

[1] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Language Process., vol.19, no.4, pp.788–798, 2011.
[2] F. Bahmaninezhad and J.H.L. Hansen, "i-vector/plda speaker recognition using support vectors with discriminant analysis," Proc. ICASSP, pp.5410–5414, 2017.
[3] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," IEEE 11th International Conference on Computer Vision, 2007, ICCV 2007, pp.1–8, 2007.
[4] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne, pp.105–111, 2018.
[5] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," 2018 IEEE Spoken Language Technology Workshop (SLT), 2018.
[6] Y. Liu, L. He, J. Liu, and M.T. Johnson, "Speaker embedding extraction with phonetic information," Interspeech 2018, pp.2247–2251, 2018.
[7] S.O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," Proc. INTERSPEECH, pp.1353–1357, 2017.
[8] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," Proc. INTERSPEECH, pp.818–822, 2016.
[9] H. Miyamoto, S. Shiota, and H. Kiya, "Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts," Proc. APSIPA Annual Summit and Conference, 2018.
[10] P.S. Nidadavolu, C.-I. Lai, J. Villalba, and N. Dehak, "Investigation on bandwidth extension for speaker recognition," Proc. INTERSPEECH, pp.1111–1115, 2018.
[11] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification," Proc. INTERSPEECH, pp.4055–4059, 2019.
[12] L.F. Gallardo, M. Wagner, and S. Möller, "I-vector speaker verification for speech degraded by narrowband and wideband channels," Speech Communication; 11. ITG Symposium, pp.1–4, 2014.
[13] K. Sriskandaraja, V. Sethu, P.N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," Proc. INTERSPEECH, pp.1710–1714, 2016.
[14] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," Proc. ICASSP, pp.6087–6091, 2014.
[15] P.N. Le, E. Ambikairajah, E.H.C. Choi, and J. Epps, "A nonuniform subband approach to speech-based cognitive load classification," Proc. ICICS, pp.1–5, 2009.
[16] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," Fifteenth annual conference of the international speech communication association, 2014.
[17] T. Thiruvaran, V. Sethu, E. Ambikairajah, and H. Li, "Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition," Electronics Letters, vol.51, no.25, pp.2149–2151, 2015.
[18] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," Proc. ICASSP, pp.5429–5433, 2018.
[19] J. Abel and T. Fingscheidt, "Sinusoidal-based lowband synthesis for artificial speech bandwidth extension," IEEE/ACM Trans. Audio, Speech, Language Process., vol.27, no.4, pp.765–776, April 2019.
[20] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," Proc. INETRSPEECH, pp.3697–3701, 2017.
[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," IEEE 2011 workshop on automatic speech recognition and understanding, 2011.
[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," Proc. ICASSP, 2018.
[23] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," Proc. INTERSPEECH, pp.999–1003, 2017.
[24] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," Proc. BIOSIG, pp.1–6, 2014.
[25] J. Gao, J. Du, and E. Chen, "Mixed-bandwidth cross-channel speech recognition via joint optimization of dnn-based bandwidth expansion and acoustic modeling," IEEE/ACM Trans. Audio, Speech, Language Process., vol.27, no.3, pp.559–571, 2019.
[26] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.1, pp.71–83, 2018.
[27] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," Proc. ICASSP, pp.4395–4399, 2015.
[28] E. Larsen, R.M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," Audio Engineering Society Convention 112, 2002.
[29] A. Nagrani, J.S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," Interspeech 2017, pp.2616–2620, 2017.
[30] J.S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," Interspeech 2018, pp.1086–1090, 2018.
[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.
[32] T. Ko, V. Peddinti, D. Povey, M.L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," Proc. ICASSP, pp.5220–5224, 2017.

Ryota Kaminishi received his B.Eng. degree from Tokyo Metropolitan University, Japan in 2017. From 2017, he has been a Master course student at Tokyo Metropolitan University. He is a member of the Acoustical Society of Japan (ASJ).

Haruna Miyamoto received her B.Eng. degree from Tokyo Metropolitan University, Japan in 2018. From 2018, she has been a Master course student at Tokyo Metropolitan University. She is a member of the Acoustical Society of Japan (ASJ).



Sayaka Shiota received the B.E., M.E. and Ph.D. degrees in intelligence and computer science, Engineering and engineering simulation from Nagoya Institute of Technology, Nagoya, Japan in 2007, 2009 and 2012, respectively. From February 2013 to March 2014, she had worked at the Institute professor. In April of 2014, she joined Tokyo Metropolitan University as an Assistant Professor. Her research interests include statistical speech recognition and speaker verification. She is a member of Acoustical Society of Japan (ASJ), IPSJ, IEICE, APSIPA, and IEEE.

Hitoshi Kiya received his B.Eng. and M.Eng. degrees from Nagaoka University of Technology, Japan, in 1980 and 1982, respectively, and his D.Eng. degree from Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University as an Assistant Professor, where he became a Full Professor in 2000. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He was/is the Chair of IEEE Signal Processing Society Japan Chapter, an Associate Editor for IEEE Trans. Image Processing, IEEE Trans. Signal Processing and IEEE Trans. Information Forensics and Security, respectively. He also serves/served as the President of IEICE Engineering Sciences Society (ESS), the Editor-in-Chief for IEICE ESS Publications, and the President-Elect of APSIPA. He is a Fellow of IEEE, IEICE and ITE.