*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification

*Ryota Kaminishi*[1], *Haruna Miyamoto*[1], *Sayaka Shiota*[1], *Hitoshi Kiya*[1]

[1]Department of Computer Science, Tokyo Metropolitan University, Japan

kaminishiryota4869@gmail.com, miyamoto-haruna@ed.tmu.ac.jp, sayaka@tmu.ac.jp, kiya@tmu.ac.jp

## Abstract

This study evaluates the effects of some non-learning blind bandwidth extension (BWE) methods on automatic speaker verification (ASV) systems based on x-vector. Recently, a non-linear bandwidth extension (N-BWE) 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 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 x-vector when the mismatches were present. 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.

**Index Terms**: Automatics speaker verification, x-vector, Non-linear bandwidth extension

## 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-based 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 NIST speaker recognition evaluation (SRE) or Speaker In the Wild (SITW) [7, 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 important factor for ASV performance. However, sampling mismatches that seriously degrade ASV performance occur between training and data evaluation because ASV systems are based on statistical machine learning frameworks. Therefore, it is difficult to unify the NB and WB databases for a ASV system. When mismatches are present, data that has a higher sampling rate is usually downsampled to a lower one [9]. 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. 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 [10–16]. 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 [10]. It has been reported that N-BWE performed well in terms of speaker individuality and root mean square log-spectral distortion (RMS-LSD). Additionally, non-learning BWE approaches are also reported [16–19] in recent years.

Although it has been reported that some ASV approaches estimate models with NB and WB mixed data [11], few studies have investigated the effects of applying non-learning BWE methods to ASV systems. This paper is focused on the non-learning BWE methods and the effects they have on x-vector-based ASV systems. For training x-vector-based ASV systems, there are three portions of dataset: for training speaker independent models, for estimating enrollment x-vector and for evaluation. Therefore, we assume two situations as sampling-rate mismatch problems. One is that the data for speaker independent models is in WB conditions, but the enrollment and the evaluation data are sampled in NB conditions. The other one is that the data for speaker independent and enrollment models are sampled in WB conditions, but the evaluation data are sampled at NB conditions. These mismatch problems are depended on applications and this problem will also face between WB and super WB conditions. Since the non-leaning BWE methods have some possibilities to relax the mismatch problems, this paper investigates there effectiveness.

To evaluate the effectiveness of the BWE methods, we carried out an 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 database.

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 section 4 illustrates our experimental setup and the results. Finally, section 5 concludes the paper.

## 2. Automatic speaker verification systems

### 2.1. X-vector

A recent ASV system based on the x-vector is another recently developed state-of-the-art system called "x-vector" [20]. Speaker individuality is represented by DNN embeddings [21]. The DNN structure is showed in Fig. 1. 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 and third



http://dx.doi.org/10.21437/Interspeech.2019-1510

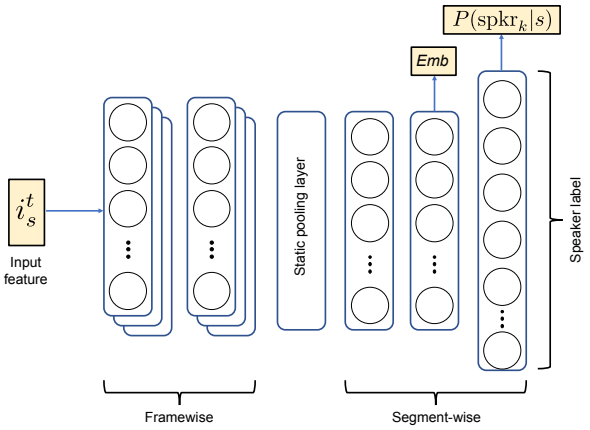


Figure 1: *DNN structure for x-vector*

layers of the DNN structure in Fig. 1 works with the framewise input features. The embedding layers (Emb) are trained with segment-wise features through the static pooling layer.

### 2.2. PLDA

For ASV back-end systems, a PLDA classifier is used [22, 23]. On PLDA-based frameworks, an extracted vector $\omega_u$ from an utterance $u$, is assumed to be an observation from a probabilistic generative model as

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

where $\Phi$ and $\Gamma$ are basis matrices that span speaker and channel subspace. $\delta$ and $\zeta_u$ express channel and speaker factors as standard Gaussian distributions. $\epsilon_u$ 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 offset in x-vector space. In Eq. (1), the probability generation model is defined as follows,

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

When the vectors of enroll speaker $\omega_1$ and test speaker $\omega_2$ are obtained, an identification score is calculated with two hypotheses, which were in the same speaker model and in the different speaker models. The PLDA-based back-end approach can reduce the acoustic fluctuation and improve ASV system performance.

## 3. Bandwidth extension methods

### 3.1. Categories of BWE methods

In the last decade, many bandwidth extension (BWE) methods have been developed [10, 11, 17, 24]. 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 [11, 24–26] thanks to

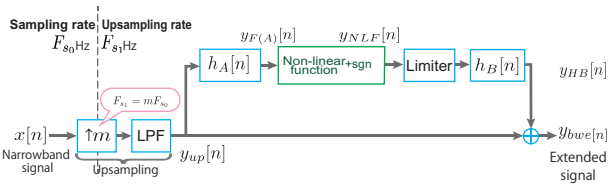


Figure 2: *Block diagram of non-linear BWE method*

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 [10, 16–18]. 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 [27]. 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].

### 3.3. Linear prediction-based analysis-synthesis

Linear prediction-based analysis-synthesis (LPAS) was developed in [17] 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

A N-BWE method has been proposed as a blind and non-learning BWE approach [10]. Fig. 2 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. A non-linear function can be used to generate harmonic components, and a general form is given by

$$y_{NLF}[n] = \text{sgn}(y_{F(A)}[n]) \cdot |y_{F(A)}[n]^{\alpha}| \times \beta, \tag{3}$$

with

$$\text{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}, \tag{4}$$

where $\alpha$ and $\beta$ are the parameters for controlling the nonlinearity, and $a$ is a real value. To control the bandwidth of $y_{FA}[n]$, the impulse response of a digital filter, $h_A[n]$ in Fig. 2, is assumed to be an all pass filter. Based on the procedure in Fig. 2, it is expected that $y_{HB}[n]$ will compensate for high-frequency losses.

### 3.5. Spectrogram comparison of each method

Figure 3 shows spectrogram examples of speech signals. First, the original signal (a) sampled at 16 kHz has frequency components from 0 kHz to 8 kHz. The upsampled signal (b) from 8

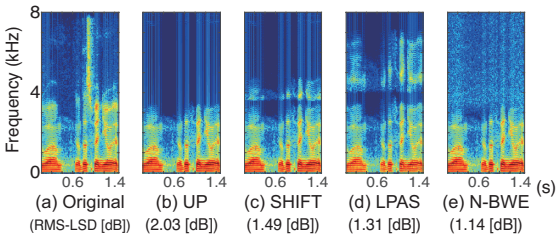


Figure 3: *Spectrogram examples of speech signals and RMS-LSD values ( $m$ = 2; $F_{S_0} = 8kHz$, $F_{S_1} = 16kHz$)*

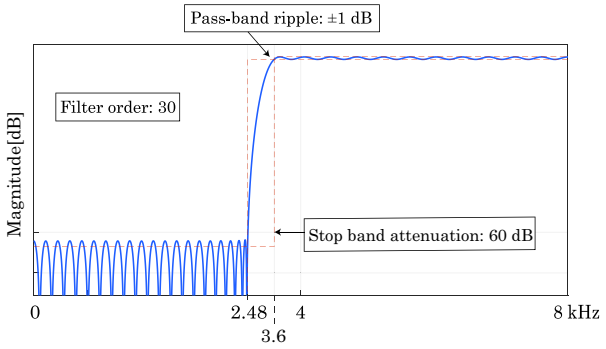


Figure 4: *Filters designed for N-BWE*

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. 3 (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. 3. 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, the N-BWE score was the lowest of all the BWE methods.

# 4. Experiments

## 4.1. Dataset description

To evaluate the effectiveness of the BWE methods, we carried out ASV experiment based on x-vector and some objective evaluations. The Kaldi toolkit [28] and a recipe for the SITW database [8] were used to construct an ASV system. The Voxceleb dataset was used to estimate the DNN and 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. 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. Two noise databases were used for data augmentation, MUSAN [31] and RIRNOISE [32]. The details can be found in [8].

Table 1: *Experimental conditions for each method*

| Scenario | Condition | Enroll | Test |
|---|---|---|---|
| Mismatch scenario (First) | (A) UP | UP | UP |
| | (B) SHIFT | SHIFT | SHIFT |
| | (C) LPAS | LPAS | LPAS |
| | (D) N-BWE | N-BWE | N-BWE |
| Mismatch scenario (Second) | (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 |

## 4.2. Experimental setup

For training x-vector-based ASV systems, there are three portions of dataset: for training speaker independent models, for estimating enrollment x-vector and for evaluation. Therefore, we assume two scenarios as sampling-rate mismatch problems. The first one is that the data for speaker independent models is in WB conditions, but the enrollment and the evaluation data are sampled in NB conditions. The second one is that the data for speaker independent and enrollment models are sampled in WB conditions, but the evaluation data are sampled at NB conditions. The EER was used as an evaluation measurement. 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).

For acoustic features, we used 30 MFCCs plus a log energy computed over a window of 25 ms with a frame shift of 20 ms. Delta and acceleration were appended to create 60 dimensional feature vectors. Although the DNNs and PLDA models were trained with Voxceleb 1 and Voxceleb 2, the training models required too much time because both databases contained over 1,000,000 utterances. The DNN was trained after reducing the number of utterances from 1,000,000 to 100,000 utterances in Voxceleb1 and Voxceleb 2 as clean data. The PLDA was trained after augmenting the clean data by using MUSAN and RIRNOISE database.

## 4.3. Comparison conditions

The following conditions were compared. The WB signals were sampled at 16 kHz and the NB signals were sampled at 8 kHz.

**(A) UP(First)**
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 (First)**
The enroll and test data was extended by SHIFT [16]. The band-pass filer was the same as [27].

**(C) LPAS (First)**
All data for enroll and test was extended by LPAS [17] from the narrowband speech sampled at 8kHz.

**(D) N-BWE (First)**
The enroll and test data was extended by using the N-BWE method [10] from the narrowband 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. 4. To

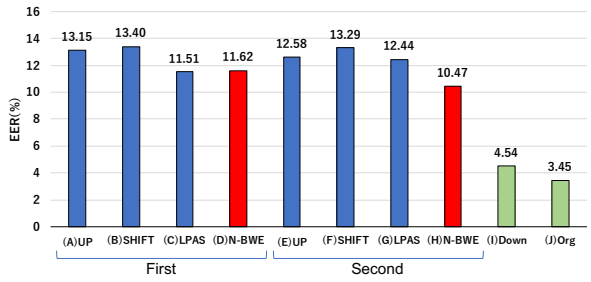


Figure 5: *EERs for each conditions on ASV systems based on x-vector (Evaluation task)*

control the nonlinearity, $\alpha$ and $\beta$ in Eq.(3) were set to 2 and 100,000, respectively.

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

**(F) SHIFT (Second)**
From (E), the test data was extended by SHIFT. The band-pass filer used [27].

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

**(H) N-BWE (Second)**
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 narrowband signal $x[n]$ in Fig. 2.

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

### 4.4. Results

Figure 5 shows the EERs on the x-vector systems for each of the conditions. Comparing the EER of (I) with that of (J), when the sampling rate mismatch was not present, the ASV performance did not changed 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 problem. The results of the mismatch scenario "First" and "Second" had a similar tendency. The EERs of (A) and (E) were high due to the missing information, although the EERs of SHIFT (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.

Figure 6 illustrates the PESQ, STOI or RMS-LSD scores for each BWE method. From the results, N-BWE obtained a lower EER, slightly higher PESQ and STOI scores, and a lower RMS-LSD value than LPAS. The tendencies of the EERs corresponded with the objective scores. From these results, N-BWE had a potential for improving the performance of ASV systems based on x-vector in sampling rate mismatch scenarios.

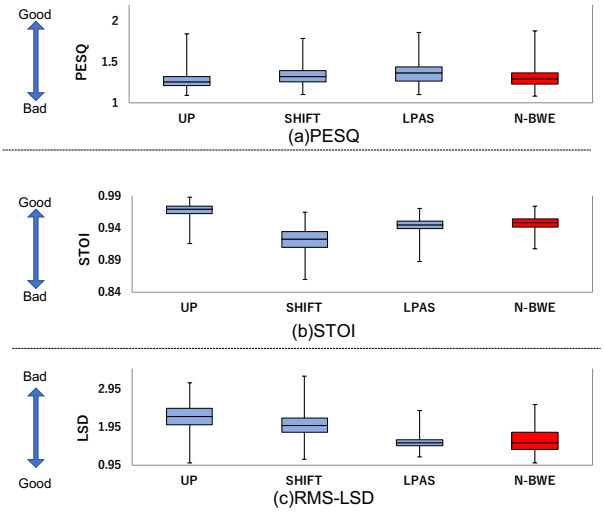


Figure 6: *Objective results for each BWE method*

## 5. Conclusion

This paper evaluated the effects of some non-learning and blind BWE methods on ASV systems based on 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 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. Additionally, since the BWE methods can use as a technique for data augmentation, the effectiveness will be evaluated.

## 6. Acknowledgement

This work was supported in part by Grant-in-Aid for Young Scientists (B), 16757733, and JSPS KAKENHI Early-Career Scientists Grant number JP19K20271.

## 7. References

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre 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] Johan Rohdin, Anna Silnova, Mireia Diez, Oldřch Plchot, Pavel Matějka, and Lukáš Burget, "End-to-end dnn based speaker recognition inspired by i-vector and plda," *in Proc. ICASSP*, pp. 4874–4878, 2018.

[3] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, 2007.

[4] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken language

recognition using x-vectors," *Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne*, 2018.

[5] Suwon Shon, Hao Tang, and James Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," *arXiv preprint arXiv:1809.04437*, 2018.

[6] Yi Liu, Liang He, Jia Liu, and Michael T Johnson, "Speaker embedding extraction with phonetic information," *arXiv preprint arXiv:1804.04862*, 2018.

[7] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," *Interspeech*, pp. 1353–1357, 2017.

[8] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The speakers in the wild (sitw) speaker recognition database.," *Interspeech*, pp. 818–822, 2016.

[9] Laura Fernández Gallardo, Michael Wagner, and Sebastian Möller, "I-vector speaker verification for speech degraded by narrowband and wideband channels," *Speech Communication; 11. ITG Symposium*, pp. 1–4, 2014.

[10] 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*, 2018.

[11] Phani Sankar Nidadavolu, Cheng-I Lai, Jesús Villalba, and Najim Dehak, "Investigation on bandwidth extension for speaker recognition," *Proc. Interspeech*, pp. 1111–1115, 2018.

[12] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech.," *INTERSPEECH*, pp. 1710–1714, 2016.

[13] H. Seo, H.G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," *in Proc. ICASSP*, pp. 6087–6091, 2014.

[14] P. N. Le, E. Ambikairajah, E. H. Choi, and J. Epps, "A nonuniform subband approach to speech-based cognitive load classification," *in Proc. ICICS*, pp. 1–5, 2009.

[15] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Fifteenth annual conference of the international speech communication association*, 2014.

[16] 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.

[17] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," *in Proc. IEEE International Conference on Acoustics, Speech and Signal*, pp. 5429–5433, 2018.

[18] J. Abel and T. Fingscheidt, "Sinusoidal-based lowband synthesis for artificial speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 765–776, April 2019.

[19] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," *Proc. Interspeech*, pp. 3697–3701, 2017.

[20] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *in Proc. ICASSP*, 2018.

[21] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech*, pp. 999–1003, 2017.

[22] Tomi Kinnunen, Lauri Juvela, Paavo Alku, and Junichi Yamagishi, "Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation," *in Proc. ICASSP*, pp. 5535–5539, 2017.

[23] Federico Alegre, Artur Janicki, and Nicholas Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," *in Proc BIOSIG*, pp. 1–6, 2014.

[24] Jianqing Gao, Jun Du, and Enhong Chen, "Mixed-bandwidth cross-channel speech recognition via joint optimization of dnn-based bandwidth expansion and acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 559–571, 2019.

[25] Johannes Abel and Tim Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018.

[26] K. Li and C. Lee, "A deep neural network approach to speech bandwidth expansion," *in Proc. ICASSP*, pp. 4395–4399, 2015.

[27] Erik Larsen, Ronald M Aarts, and Michael Danessis, "Efficient high-frequency bandwidth extension of music and speech," *Audio Engineering Society Convention 112*, 2002.

[28] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[29] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[31] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *in Proc. ICASSP*, pp. 5220–5224, 2017.