Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts

Haruna Miyamoto*, Sayaka Shiota* and Hitoshi Kiya*
* Tokyo Metropolitan University, Faculty of System Design, Department of Computer Science, Japan
E-mail: miyamoto-haruna@ed.tmu.ac.jp

***Abstract*—This paper has three aims that are to point out that signals generated by conventional BWE methods generally include some aliasing artifacts, to propose a novel bandwidth extension (BWE) method considering the effects of aliasing artifacts, and to apply various BWE methods to speaker verification to evaluate the effectiveness of the BWE ones. Study on BWE methods using non-linear functions has a long history started from for analog signal processing, but conventional BWE ones have never considered the influence of aliasing artifacts caused by the band limitation that digital signals have. This paper is among the first to point out that discrete-time signals generated by BWE methods generally include some aliasing artifacts due to the band limitation to be decided according to the sampling frequency. Next, a new non-linear artificial BWE method, considering of aliasing artifacts, is proposed. Moreover, to evaluate the proposed framework, speaker verification experiments and objective tests are conducted. Experiment results show that speech signals extended by the proposed framework provide the error reduction of 46.2%, compared with a typical conventional method. Additionally, the generated speech signals are evaluated by using three object measures: PESQ, RMS-LSD and STOI. It is also implied that equal error rates in speaker verification tasks have a closest relation with RMS-LSD in with another measures.**

I. Introduction

In communication networks, bandwidth of transmitted speeches is typically limited to 3.4 kHz. The limited narrowband speech signals lead to serious degradations of speech quality, naturalness, and speaker individuality. It is also well-known that performances of statistical-based machine leaning systems such as speech recognition, speech synthesis and speaker verification are strongly affected by the band-limitation. To improve the degradations, various bandwidth extension (BWE) methods have been studied so far, where BWE methods are regarded as one of techniques to restore high frequency losses caused by the bandwidth limitation, sometimes referred to as super resolution ones. This paper aims to point out that signals generated by conventional BWE methods generally include some aliasing artifacts, then to propose a new BWE method considering aliasing artifacts, and finally to evaluate the effectiveness of the proposed BWE method in speaker verification experiments.

BWE methods have been conducted in various research fields such as speech processing, image and video one, and acoustics signal one. Many different approaches to BWE have been reported later and that can be categorized as either blind or non-blind. Non-blind methods restore missing frequency components from auxiliary high-frequency (HF) side information which is encoded into a data stream together with low-frequency (LF) components. In contrast, blind methods estimate missing HF components using only the LF components. Additionally, they can be broadly classified into two types according to a difference in computational approaches: non-leaning-based type, to which the proposed method corresponds, and leaning-based type. This paper focuses on the former one, that a light-weight computational cost. Non-linear BWE methods including rule-based spectrum folding [1]–[5] and mapping approaches [6]–[8] are in the former type, while statistical approaches using Gaussian mixture models (GMMs) [9], [10], hidden Markov models (HMMs) [11]–[15] or neural networks [16]–[18] belong to the latter one. In the BWE methods, the quality extended of signals has been mostly evaluated according to objective measures such as signal-to-noise ratio (SNR) and log spectral distance (LSD). However, it has been reported that the measures are not always suitable for performance evaluation of statistical machine learning-based systems such as speech recognition [19]. In [20], it has been also shown that some non-leaning-based methods outperform GMM-based methods in terms of LSD values, but the performance evaluation of statistical machine learning-based systems have not been carried out.

Because of such a situation, this paper discusses non-linear BEW methods and its performance evaluation. Study on BWE methods using non-linear functions has started from for analog signal processing [21], however conventional BWE ones have never considered the influence of aliasing artifacts caused by the band limitation that digital signals have. In this paper, it is pointed out that discrete-time signals generated by BWE methods include some aliasing artifacts due to the band limitation. BWE methods can be categorized into according to various criteria. One category is a blind or non-blind method [22]. The second one is whether belongs to source-filter or non-source filter model [23]. The other ones are a time-domain or transform-domain BWE algorithm [24] and an approach based on decoding or information hiding [25]. In these cases, the proposed method is regarded as a blind, non-source filter, transform-domain, and decoding approach. In order to evaluate the proposed method, speaker verification experiments with a GMM-UBM system and objective tests are conducted. From the results, the proposed BWE methods from 8 kHz to 16 kHz have an error reduction of 46.2% compared

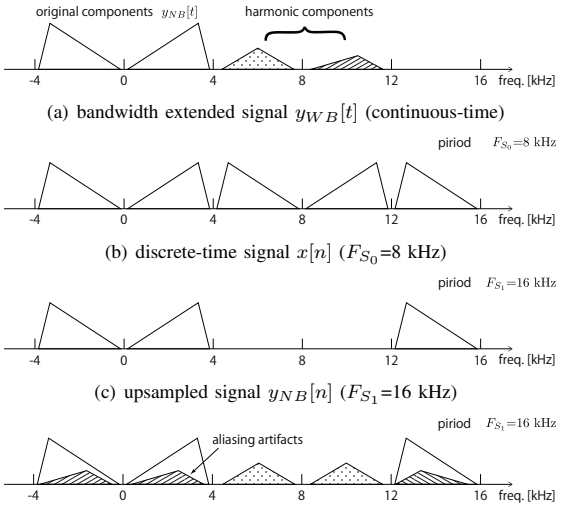


Fig. 1: non-linear BWE ($K = 2$)

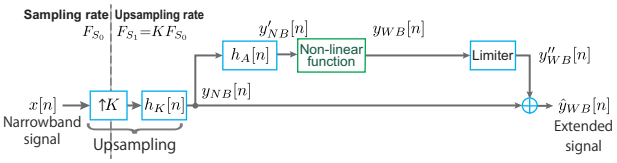


Fig. 2: Block diagram of conventional BWE

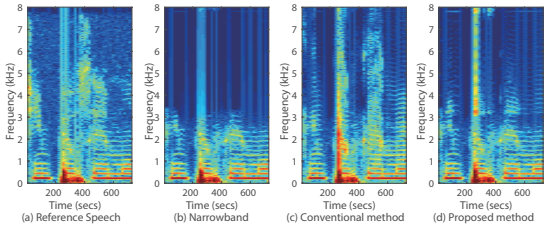


Fig. 3: Spectrogram examples of speech signals ($F_{S_0} = 8$ kHz, $F_{S_1} = 16$ kHz)

with the conventional methods [26], [27]. Additionally, the generated speeches are evaluated with perceptual evaluation of speech quality (PESQ), short-time objective intelligibility measure (STOI) and root mean square log spectral distortion (RMS-LSD). These results are shown that the proposed method outperforms conventional BWE methods in term of RMS-LSD as well as equal error rate (EER) in the speech verification tasks.

II. Preparation

A. Non-linear bandwidth extension

Conventional non-linear BWE methods are summarized, here. Firstly, let $y_{NB}[t]$ be a continuous-time signal with a narrowband width as shown in Fig. 1 (a), where t is used to denote the continuous-time variable. To produce a signal with a wider bandwidth $y_{WB}[t]$, the use of a non-linear function $f_N(\cdot)$ has been proposed as, in [21],

$$y_{WB}[t] = f_N(y_{NB}[t]). \tag{1}$$

Figure 1 (a) illustrates the relationship between $y_{NB}[t]$ and $y_{WB}[t]$ in the frequency domain. As shown in the figure, some harmonic components are generated from $y_{NB}[t]$ by using the non-linear function, so $y_{WB}[t]$ has not only the original frequency components but also the harmonic ones.

This principle has been extended to discrete-time signals [20], [28]–[33]. In [30]–[33], a non-linear BWE method has been studied as an efficient scheme for image super-resolution. Figure 2 shows the block diagram of the procedure in [30]–[33]. In Fig. 2, $x[n]$ is a discrete-time signal with the same bandwidth as $y_{NB}[t]$ in Fig. 1, where n is for the discrete-time variable. By using an upsampler with an integer factor K and a linear digital filter $h_K[n]$, an upsampled signal $y_{NB}[n]$ is generated as shown in Fig. 1 (c). Note that the spectrum of $y_{NB}[n]$ is periodic and its period is given by the sampling frequency, F_{S_1}. In addition, $y_{NB}[n]$ has no harmonic components. A non-linear function can be used to generate harmonic components as well as for continuous-time signals. In [30]–[33], a general form of non-linear functions is given by,

$$y_{WB}[n] = sgn(y'_{NB}[n]) \cdot |y'_{NB}[n]^{\alpha}| \times \beta, \tag{2}$$

with

$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}, \tag{3}$$

where α and β are parameters to control the non-linearity, and a is a real value. These functions have been successfully applied to image super-resolution. The delay in Fig. 2 depends on filter order.

However, we have to take special care to the periodicity that the spectrum of discrete-time signals has. In other words, the bandwidth of $y_{NB}[n]$ is limited to $F_{S_1}/2$. Figure 1 (d) demonstrates the effect of the band limitation, where the inpulse responce of a digital filter, $h_A[n]$ in Fig. 2 is assumed as

$$h_A[n] = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}, \tag{4}$$

for simplifying the discussion. From the figure, $y_{NB}[n]$ includes some aliasing artifacts, although it also includes some useful harmonic components.

One of our aims is to propose a new BWE method considering the aliasing artifacts. This paper is among the first to point out the effect of the artifacts in the BWE.

B. Necessity for speech signals

In Fig. 3, some spectrogram examples of speech signals are demonstrated in terms of the difference in high frequency

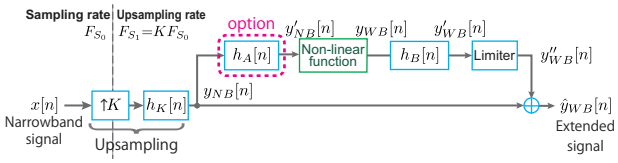
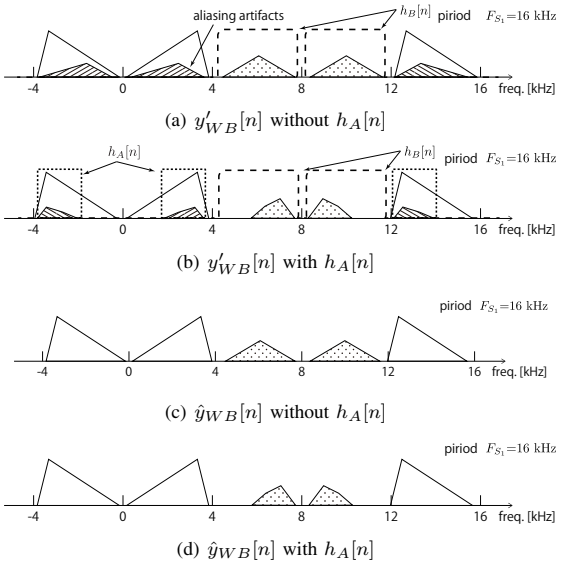


Fig. 4: Block diagram of proposed BWE



(a) $y'_{WB}[n]$ without $h_A[n]$

(b) $y'_{WB}[n]$ with $h_A[n]$

(c) $\hat{y}_{WB}[n]$ without $h_A[n]$

(d) $\hat{y}_{WB}[n]$ with $h_A[n]$

Fig. 5: Proposed BWE ($K = 2$)

components, in which the reference signal in (a) has 8 kHz bandwidth, (b) is a speech with 4 kHz bandwidth produced from the reference one, (c) is a wideband speech generated from the speech in (b), by using the conventional BWE method [30], and (d) is a wideband speech generated with the proposed BWE method. From these examples, both of the BWE methods in Figs. (c) and (d) can generate some harmonics components in the high frequency band (4 kHz–8 kHz). Furthermore, comparing the lower frequency components in Fig. 3 (c) with those in Fig. 3 (d), it is confirmed that the proposed method allows us to reduce aliasing artifacts. This paper aims to propose a new BWE method considering aliasing artifacts, and to evaluate the effectiveness of the method.

It is well known that band-limited speech signals degrade intelligibility and speaker's characteristics. Moreover, the performances of statistical-based machine leaning systems such as speech recognition, speech synthesis and speaker verification are strongly affected by the band-limitation. In the telephone communication systems, narrowband signals, up to 3.4 kHz, have been used. Meanwhile, recently, wideband or super-wideband communications, which use a wider bandwidth than 3.4 kHz, also come to be used, so there are various speech signals with different bandwidths in actual communication systems. As a result, the signals mixed with various bandwidths make statistical-based machine leaning systems more complex. Because of such situations, BWE methods are required to improve the situations.

III. Proposed BWE framework

A. BWE considering aliasing artifacts

To reduce the aliasing artifacts descried in 2.1, a new BWE method is proposed. Figure 4 is the block diagram of the proposed method. The difference between Fig. 2 and Fig. 4 is that there are two filters, i.e. $h_A[n]$ and $h_B[n]$ in Fig. 4. If the impulse response $h_A[n]$ meets eq. (4), $h_A[n]$ dose not provide any operation.

Figure 5 illustrates the difference in the frequency domain. As shown in Figs. 5 (a) and (c), the use of the filter $h_B[n]$ enables us to delete the aliasing artifacts that overlap the original components of a signal. Meanwhile, the filter $h_A[n]$ plays a role in the control of harmonic components due to filtering the original components.

The limiter in Fig. 4 is given by, as well as in [30].

$$y'_{WB}[n] = \begin{cases} y_{WB}[n], & y_{WB}[n] \leq T_h \\ M, & y_{WB}[n] > T_h \end{cases}, \tag{5}$$

where T_h is a threshold value and M is a constant value. Based on the procedure in Fig. 4, it is expected that $\hat{y}_{WB}[n]$ does not include aliasing artifacts, although it includes the original components.

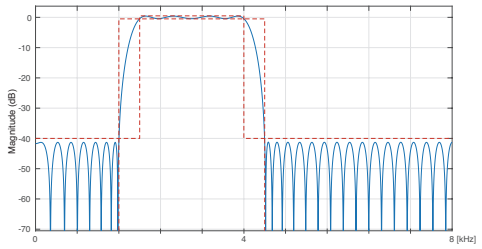
B. Filter specification

Two filters, i.e. $h_A[n]$ and $h_B[n]$ in Fig. 4 are explained here in more detail. In the proposed BWE method, $h_A[n]$ and $h_B[n]$ are designed as band-pass FIR (Finite Impulse Response) filters with the specifications given in Fig. 6, as well as in [27], although the conventional method [30] uses a high-pass filter as $h_A[n]$. The quality of $\hat{y}_{WB}[n]$ slightly depends on the specifications of filters used for the BWE method, so almost the same specifications as those of the conventional BWE one [27] are used to fairly evaluate the effect of aliasing artifacts in this paper. In the experiments, the filters in Fig. 6 are used as $h_A[n]$ and $h_B[n]$.

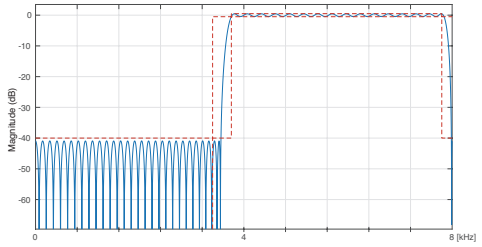
C. Application to speaker verification

So far, BWE methods have been evaluated based on some objective measures such as mean opinion score (MOS) [34] and PESQ [35], STOI [36] and RMS-LSD [37]. However, it has been reported that the measures are not always suitable for the performance evaluation of statistical machine learning-based systems [19]. In this paper, to evaluate the proposed BWE method, GMM-universal background model (GMM-UBM)-based speaker verification systems are used as one of typical speaker verification systems [38]. The flow of the verification system with BWE is illustrated in Fig. 7.

In the enrollment part, databases are used to estimate an UBM as the speaker independent model. The UBM is represented as a GMM, and the speaker dependent model λ_A is estimated from the UBM and the feature vector of



(a) Band-pass filter $h_A[n]$ with pass-band cutoff frequencies $F_{p1} = 2$ kHz and $F_{p2} = 4$ kHz, stop-band frequencies $F_{s1} = 1.6$ kHz and $F_{s2} = 4.4$ kHz, a pass-band ripple of ± 1 dB, a stop-band attenuation of 40 dB, and a filter order of 64.



(b) Band-pass filter $h_B[n]$ with pass-band cutoff frequencies $F_{p1} = 3.76$ kHz and $F_{p2} = 7.84$ kHz, stop-band frequencies $F_{s1} = 3.44$ kHz and $F_{s2} = 8$ kHz, a pass-band ripple of ± 1 dB, a stop-band attenuation of 40 dB, and a filter order of 120.

Fig. 6: Filters designed for the proposed BWE

an enrollment speaker A by using the maximum a posterior (MAP) adaptation. In Fig. 7, the database consists of signals sampled at 8 kHz, and then BWE methods are adopted to estimate signals sampled at 16 kHz from the signals of the database. The speech samples of an enrollment speaker are also extended to signals sampled at 16 kHz as well.

In the verification part, BWE methods are applied to a query speech sample, and the verification score $S(\boldsymbol{X}, \lambda_A)$ is calculated between the query feature vectors $\mathbf{X}$ and the enrollment speaker model λ_A. According to a threshold θ, the query is decided to accept or not. Under this flow,this paper aims to discuss the relationship between the subjective measurements and the performance of the speaker verification system with the conventional BWE methods or the proposed one.

IV. Experiment

To evaluate the effectiveness of the proposed method, speaker verification experiments and objective tests were performed.

A. Experimental condition

For speaker verification systems, the GMM-UBM-based framework was conducted as Fig. 7. Table I summarizes

TABLE I: Experimental conditions for GMM-UBM systems

Database (UBM)	JNAS [39](female only) 16 kHz sampling
Training data (UBM)	23,657 sentences
Database (Speaker dependent (SD) model)	VLD database [40] (Headset microphone) 48 kHz sampling
# of Speaker	17 (female only)
Training data (SD)	70 sentences / speaker (Total 1190 sentences)
Test data	30 sentences / speaker (Total 510 sentences)
# of mixtures	1,024
Frame length	25 msec
Frame shift	10 msec
Feature	MFCC 19 order + Δ + $\Delta\Delta$

experimental conditions for constructing the GMM-UBM-based speaker verification systems with the BWE methods. To estimate an UBM as the speaker independent model, Japanese Newspaper Article Sentences (JNAS) database [39], which contains over 150 female speakers, was utilized. As an enrollment database, Voice Liveness Detection (VLD) database [40] with 17 female speakers was used. The JNAS and VLD databases were recorded at 16 kHz and 48 kHz, respectively. Therefore, the signals in the VLD database were downsampled from 48 kHz to 16 kHz in order to generate reference speech signals. A speaker dependent (SD) model was estimated with 70 sentences for each enrollment speaker by using the MAP adaptation. Besides, other 30 sentences per enrollment speaker were used for test set. In the speaker verification systems with the BWE methods, every speech sample was downsampled from 16 kHz to 8 kHz to be used as the narrowband signals $x[n]$. Then, each BWE method was applied to the narrowband signals. The following methods were compared, under T_h=0.001 and the use of MATLAB R2017a.

(A) UP
: In the method " (A) UP," all speech data for the UBM, SD models and the test set were generated from the narrowband data sampled at 8kHz by the upsampling operation with $K = 2$, as $y_{NB}[n]$ in Fig 1 (c). Note that the speech data did not include any harmonic components in the high frequency components.

(B) Conv. I
: All speech data were generated from the data given in " (A) UP" by applying the conventional BWE method [26], [27]. In Conv. I, $h_A[n]$ and $h_B[n]$ were used as band-pass filters in Fig. 6.

(C) Conv. II
: All speech data were generated from the data given in " (A) UP" by applying the conventional BWE method [30], where α and β in Eq. (1) were experimentally set to 1.8 and 100, respectively. $h_A[n]$

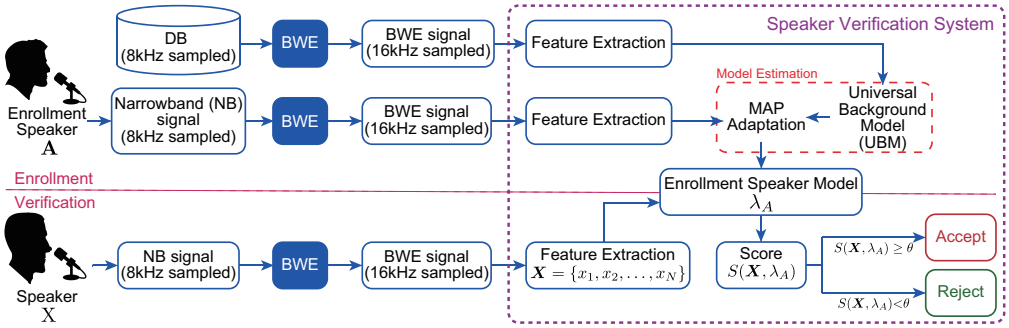


Fig. 7: GMM-UBM Speaker Verification system with BWE methods

in Fig.2 was a high-pass filter with a stop-band frequency $F_s = 2.4$ kHz, a pass-band cutoff frequency $F_p = 3.6$ kHz, a pass-band ripple of ± 1 dB, a stop-band attenuation of 60 dB, and a filter order of 28.

(D) LPAS
All speech data were generated from the narrowband data sampled 8 kHz by applying the Linear Prediction based Analysis-Synthesis (LPAS) method, which has been proposed as one of blind and non-learning BWE method [5].

(E) Proposed I
All speech data were generated from the data given in " (A) UP" by the proposed method. $h_A[n]$ meets Eq. (4). α and β in Eq. (1) were set to 1.8 and 100, respectively.

(F) Proposed II
All speech data were generated from the data given in " (A) UP" by the proposed method, with the two band-pass filters in Fig. 6, where α and β were to 1.5 and 100, respectively.

(G) 8k
The narrowband data sampled at 8 kHz were used for all data.

(H) 16k
The reference data sampled at 16 kHz were used for all data.

The evaluation of speaker verification systems were carried out with equal error rates (EERs).

For the objective tests of speech quality, PESQ, STOI and RMS-LSD were used. RMS-LSD stand for the log spectral distance between two signals given by,

$$D = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{N}\sum_{i=0}^{N-1}\left|\log_{10} A_k(i) - \log_{10}\hat{A}_k(i)\right|^2}, \quad (6)$$

where $A_k(i)$ and $\hat{A}_k(i)$ were power spectrums in k frame of the reference speech and the generated speech by the BWE methods. N and k indicated the frame length and the frame number, respectively. A low RMS-LSD value meant the generated speech was close to the reference one. The reference and narrowband signals were prepared by the same manner as the speaker verification tasks. For the tests, the number of the pairs of the reference and narrowband signals was 1,700.

B. Experimental results

Figure 8 shows the EERs of the speaker verification systems with each method. The BWE methods were carried out under $F_{S_0} = 8$ kHz, $F_{S_1} = 16$ kHz and $K = 2$ as shown in Fig. 4. From the comparison of (G) and (H), it was recognized that the performance of the speaker verification systems was considerably degraded due to the effect of the narrowband limitation. The EER of (A) was lower than that of (G), even though both of them had the same bandwidth. This is because that the frequency response of filter banks in the extraction process of mel-frequency cepstral coefficients (MFCCs) depended on the sampling frequencies of input signals. Among the conventional methods (B), (C) and (D), while the EERs of (B) and (C) were worse than that of (A), (D) obtained the lower EER than (A). It indicated that the non-linear-based methods (B) and (C) were suffered from the aliasing artifacts, even though these methods were able to generate the high frequency components. The EERs of the proposed methods (E) and (F) were lower than that of (D). This result denoted that (D), (E) and (F), which contains the high-pass filter in the latter part of each BWE process were able to relax the aliasing artifacts. Furthermore, the performance of (E) and (F) were better than that of (D) due to the effective harmonics generation.

Figures 9 illustrates the objective measurement results with box plots. The box plots were drawn to visualize the distribution of scores. The top and bottom sides of the box meant the upper and lower quartiles of all results. The center quartile (median) was located inside the box as a bar. Both whiskers of the upper and lower sides represented the maximum and minimum values of the distribution.

Figure 9 (a) shows PESQ scores of each method. The PESQ score ranged from of 0 (bad) to 4.5 (best). At Fig. 9 (a), the scores of the conventional methods (B) and (C) were almost the same as that of (A) as well as the EER results. However, the scores of the proposed methods (E) and (F) were worse than those of (A), (B) and (C). The PESQ algorithm has been developed to predict the MOS score, which was a subjective measurement by comparing the reference speech with the

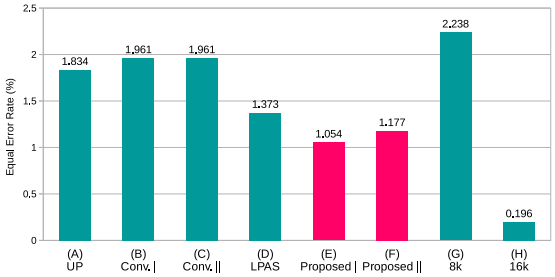
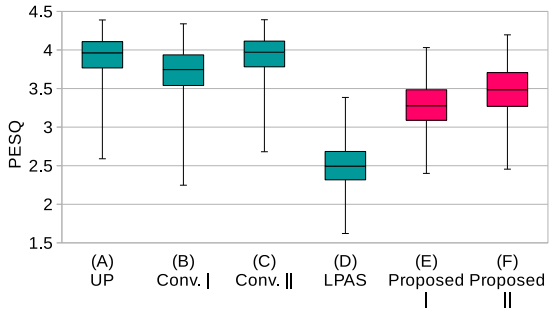
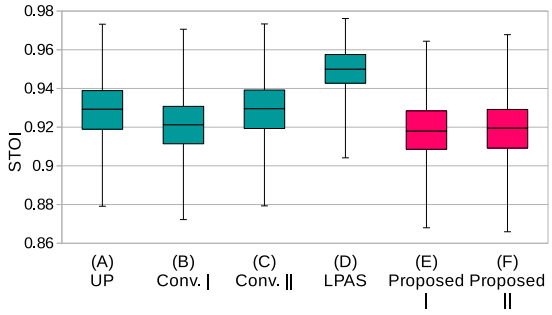


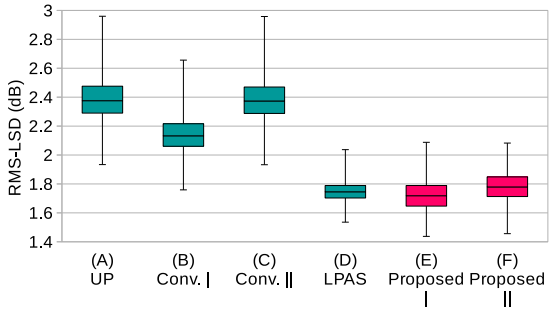
Fig. 8: Speaker verification results ($F_{S_0} = 8$ kHz, $F_{S_1} = 16$ kHz)



(a) PESQ (Reference: 16 kHzsampling data)



(b) STOI (Reference: 16 kHzsampling data)



(c) RMS-LSD (Reference: 16 kHz sampling data)

Fig. 9: Objective evaluation results

degraded one. Since the proposed methods simply generated some harmonic components from narrowband signals with a non-linear function, the proposed methods did not guarantee the improvement of the naturalness.

The second measurement is STOI, which based on a correlation coefficient between the temporal envelopes of the clean and degraded speeches, in short-time, overlapping segments. The STOI value ranged from of 0.0 to 1.0. The tendency of Fig. 9 (b) was almost the same as Fig. 9 (a) because STOI was one the measurement for the naturalness. Only the tendency between the PESQ and STOI scores of (D) was different.

Figure 9 (c) shows RMS-LSD values for each method. Even though (B) and (C) used the same non-linear function as the proposed method, both of (E) and (F) obtained lower values than (B). From this result, it indicated that the proposed methods and LPAS were able to relaxed the aliasing artifacts in the lower bandwidth and led the better RMS-LSD values as well as the EER results. From these results, it was also implied that EERs in speaker verification tasks had a closest relation with RMS-LSD in with another measures.

V. Conclusions

This paper pointed out three matters. The first one was that signals generated by the conventional non-linear-based BWE methods included some aliasing artifacts. Next, the novel non-linear BWE method was proposed to avoid the aliasing artifacts. To evaluate the proposed methods, some experiments using the GMM-UBM speaker verification systems and the objective tests were carried out. The proposed method outperformed the conventional methods in terms of both of the EER and RMS-LSD results, although it did not in terms of PESQ and STOI scores. From these results, it was also implied that EERs in speaker verification tasks had a closest relation with RMS-LSD in with another measures.

In future work, the proposed method will be evaluated with the practical communication scheme ITU-T G712 [41], and will be compared under other machine learning systems and other languages. Also, there is a possibility that the proposed method can be extended to SWBE method.

Acknowledgment

This work was supported in part by Grant-in-Aid for Young Scientists (B), 16757733.

References

[1] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," IEEE Trans. Audio, Speech, and Language. Process., vol.16, no.6, pp.1124–1137, 2008.

[2] K. Sriskandaraja, P.N.L. V. Sethu, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," in Proc. INTERSPEECH 2016, pp.1710–1714, 2016.

[3] 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, pp.2149–2151, 2015.

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

[5] 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.1–5, 2018.

[6] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech based on linear mapping," Electronics and Communications in Japan (Part II:Electronics), vol.85, no.8, pp.44–53, 2002.

[7] J. Epps and W.H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in Proc. IEEE Workshop on Speech Coding 1999, pp.174–176, 1999.

[8] N. Enbom and W.B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in Proc. IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No.99EX351), pp.171–173, 1999.

[9] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with gmm," in Proc. SLP, vol.2007, pp.63–68, 2007.

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

[11] J. Han, G.J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in Proc. IEEE Workshop on Machine Learning for Signal Process., pp.1–6, 2012.

[12] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," in IEEE Workshop on Speech Coding, pp.133–135, 2000.

[13] D.O. Y. M. Cheng and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," IEEE Trans. Speech and Audio. Process., vol.2, no.4, pp.544–548, 1994.

[14] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in Proc. ICASSP 2003, vol.1, pp.680–683, 2003.

[15] G. Song and P. Martynovich, "A study of hmm-based bandwidth extension of speech signals," Signal Processing, vol.89, no.10, pp.2036–2044, 2009.

[16] Y. Tachioka and J. Ishii, "Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition," Acoustical Science and Technology, vol.37, no.6, pp.319–321, 2016.

[17] S. Li, S. Villette, P. Ramadas, and D.J. Sinder, "Speech bandwidth extension using generative adversarial networks," in Proc. IEEE, Speech Enhancement and Recognition, pp.5029–5033, 2018.

[18] K. Schmidt and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in Proc. IEEE, Speech Enhancement, pp.5444–5448, 2018.

[19] L.F. Gallardo, S. Moller, and J. Beerends, "Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores," in Proc. INTERSPEECH 2017, pp.2939–2943, 2017.

[20] S. Asawa, Y. Sugiura, and T. Shimamura, "Voiceless consonant detection and artificial bandwidth extension of narrow band speech," IEICE Technical Report, vol.117, no.516, pp.231–234, 2018.

[21] C. Un and D. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," IEEE Trans. on Comm., vol.23, no.12, pp.1466–1474, 1975.

[22] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in Proc. EuroSpeech, vol.3, pp.1643–1646, 1997.

[23] J. Peter and P. Vary, "On artificial bandwidth extension of telephone speech," Signal Processing, pp.1707–1719, 2003.

[24] K. Tsuiino and K. Kikuiri, "Low-complexity bandwidth extension in mdct domain for low-bitrate speech coding," in Proc. ICASSP, pp.4145–4148, 2009.

[25] G. Bernd, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in Proc. INTERSPEECH, pp.1497–1500, 2005.

[26] E. Larsen, R.M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," 112th AES Convention, vol.23, no.5627, pp.1–5, 2002.

[27] 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.

[28] R. Nakanishi, S. Shiota, and H. Kiya, "Non-linear artificial bandwidth extension of narrowband speech for speaker veri cation," IPSJ SIG Technical Report, no.4, pp.1–6.

[29] R. Aarts, E. Larsen, and D. Schobben, "Improving perceived bass and reconstruction of high frequencies for band limited signals," in Proc. IEEE Workshop Model Based Processing and Coding of Audio, pp.52–71, 2002.

[30] S. Gohshi and I. Echizen, "Limitations of super resolution image reconstruction and how to overcome them for a single image," in Proc. SIGMAP 2013, pp.71–78, 2013.

[31] S. Gohshi, "Limitation of super resolution image reconstruction for video," pp.217–221, June 2013.

[32] M. Sugie and S. Gohshi, "Performance verification of super-resolution image reconstruction," pp.547–552, 2013.

[33] C. Mori, K. Tanioka, and S. Gohshi, "Super resolution image reconstruction and imaging device," pp.588–593, 2016.

[34] M.D. Polkosky and J.R. Lewis, "Expanding the mos: Development and psychometric evaluation of the mos-r and mos-x.," International Journal of Speech Technology, no.2, pp.161–182, 2003.

[35] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation, vol.862, 2001.

[36] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speach, Language. Process., vol.19, no.7, pp.2125–2136, 2011.

[37] R.M. Gray, A. Buzo, A.G. Jr, and Y. Matsuyama, "Distortion measures for speech processing," IEEE Trans. Acoustics, Speech and Signal. Process., vol.28, no.4, pp.367–376, 1980.

[38] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," Diigital Signal Processing, vol.10, pp.19–41, 2000.

[39] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T.Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus," in Proc. ICSLP 98, pp.3261–3264, 1998.

[40] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in Proc. INTERSPEECH 2015, pp.239–243, 2015.

[41] R.G. ITU-T, "Transmission performance characteristics for pulse code modulation channels," 1996.