Investigation on latency issues and objective measurements of non-linear blind bandwidth extension

Haruna Miyamoto
Tokyo Metropolitan University
Tokyo, Japan
miyamoto-haruna@ed.tmu.ac.jp

Sayaka Shiota
Tokyo Metropolitan University
Tokyo, Japan
sayaka@tmu.ac.jp

Hitoshi Kiya
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

***Abstract*—This paper investigates the algorithmic latency of non-learning blind bandwidth extension (BWE) approaches, and evaluates their quality with objective measurements. BWE methods are regarded as methods for restoring high-frequency losses caused by band limits. To satisfy the restriction and provide high-quality sound, low-latency BWE techniques are required. This paper focuses on the non-learning blind BWE methods, such as spectral shifting-based approaches and non-linear function based BWE (N-BWE). The results proved that the N-BWE method can be performed with low-latency.**

***Index Terms*—artificial bandwidth extension, non-linear function, objective measurement, latency issues**

I. Introduction

This paper investigates the algorithmic latency of non-learning blind bandwidth extension (BWE) approaches, and evaluates their quality with objective measurements. BWE methods are regarded as methods for restoring high-frequency losses caused by band limits. For some real-time applications, such as video calls, BWE methods are one of the most important methods for improving speech quality. In these applications, users receive visual and sound information simultaneously. Therefore, the time lag between both types of information is strictly restricted for natural conversations. To satisfy the restriction and provide high-quality sound, low-latency BWE techniques are required. The learning BWE methods proposed so far has high speech quality but high-latency [1], [2]. Its latency has not been discussed. Therefore, this paper focuses on the non-learning blind BWE methods, such as spectral shifting-based approaches [3]–[5] and non-linear function based BWE (N-BWE) [6], [7]. In the experiments, the algorithmic latencies of the BWE methods were calculated. The results proved that the N-BWE method can be performed with low-latency.

II. Motivation

Recently, since video calls or similar real-time applications are usually used, low-latency and high-quality techniques for speech signal processing are required. When there is a time lag between video and sound information, users perceive this as an echo, and it makes them feel uncomfortable irritated. It has been reported that users feel uncomfortable when differences between visual and sound information exceed 10 ms, and even a difference of 3–5 ms can be noticed [8]. Therefore, computational and algorithmic latencies are important issues to discuss. There are many kinds of BWE methods which can be categorized into blind or non-blind and non-learning or learning. In this paper, algorithmic latency of the non-learning blind BWE methods is focused on.

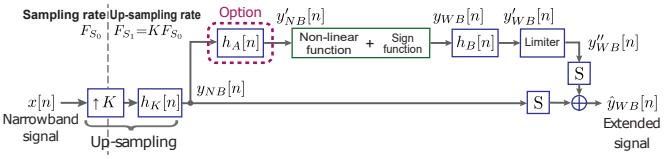


Fig. 1: Block diagram of N-BWE method

III. Non-linear BWE extension

A N-BWE method has been proposed as a blind and non-learning BWE approach [6]. Figure 1 shows a block diagram of the N-BWE method. By using basic upsampling, an upsampled signal $y_{NB}[n]$ is generated, where n is a discrete-time variable. $y_{NB}[n]$ has no harmonic components over $F_{S_0}/2$ kHz. A non-linear function can be used to generate harmonic components, and a general form of a non-linear function is given by

$$y_{WB}[n] = \mathrm{sgn}(y'_{NB}[n]) \cdot |y'_{NB}[n]^{\alpha}| \times \beta, \tag{1}$$

with

$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}, \tag{2}$$

where α and β are the parameters for controlling the nonlinearity, and a is a real value.

$$y''_{WB}[n] = \begin{cases} y'_{WB}[n], & y'_{WB}[n] \le T_h \\ M, & y'_{WB}[n] > T_h \end{cases}, \tag{3}$$

where T_h is a threshold value, and M is a constant value. S in Fig. 1 compensates for delays introduced by the different

978-1-7281-3575-5/19/$31.00 ©2019 IEEE

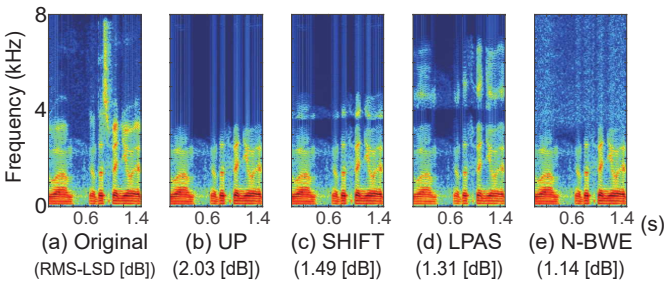
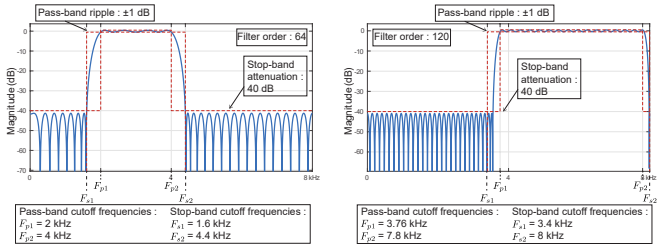


Fig. 2: Spectrogram examples of speech signals ($K = 2$, $F_{S_0} = 8$ kHz, $F_{S_1} = 16$ kHz)



(a) Band-pass filter $h_A[n]$ (b) Band-pass filter $h_B[n]$

Fig. 3: Filters designed for N-BWE I and II

processes involved in the generation of LF and HF components. On the basis of procedure in Fig. 1, it is expected that $y_{WB}[n]$, will compensate for high-frequency losses.

It has been reported that the N-BWE method provides high performance in terms of speaker individuality and root mean square log spectral distortion (RMS-LSD [9]) [7]. However, its latency has not been discussed.

A. Examples of each BWE method

Figure 2 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 kHz to 16 kHz contains only low frequency components under 4 kHz. Signals (c) and (d) were generated by shift-based BWE methods (SHIFT [3], [4] and LPAS [5]), and signal (e) was generated by N-BWE. As these examples show, the BWE methods in Figs. 2 (c), (d) and (e) can generate harmonic components in high-frequency components. The RMS-LSD scores are also shown in Fig. 2. 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 RMS-LSD score of N-BWE was the lowest of all the BWE methods.

IV. Experiment

In this section, the algorithmic latencies of each BWE method were calculated.

A. Experimental condition

All speech samples used for the experiments were collected from the Speaker In The Wild (SITW) database [10]. The SITW database consists of 4841 utterances sampled at 16 kHz from 299 speakers. The following conditions were compared.

(A) UP
: All data was simply upsampled. Note that the speech samples did not include any harmonic components in the high-frequency components.

(B) SHIFT
: All data was extended by SHIFT [3]. The band-pass filer was the same as [4].

(C) LPAS
: All data was extended by LPAS [5] from the speech sampled at 8kHz.

(D) N-BWE I
: All data was extended by using the N-BWE method [6] from the speech sampled at 8kHz. The optional filter $h_A[n]$ was defined as the all pass filter, and the filter $h_B[n]$ was band-pass filter in Figure 3 (b). To control the nonlinearity, α and β in Eq.(1) were set to 1.8 and 100, respectively.

(E) N-BWE II
: All data was extended by using the N-BWE method [6] from the speech sampled at 8kHz. The optional filter $h_A[n]$ and $h_B[n]$ were used as band-pass filters in Figs. 3 (a) and (b). To control the nonlinearity, α and β in Eq.(1) were set to 1.5 and 100, respectively.

The design of filters $h_A[n]$ and $h_B[n]$ depended only on interpolator K, and the parameters α and β also depended only on interpolator K.

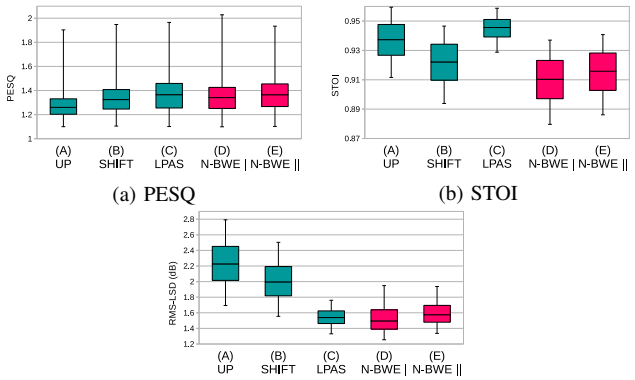
TABLE I: Algorithmic latency of each BWE method

Compared method	Latency (ms)
(A) UP	0.068
(B) SHIFT	0.643
(C) LPAS	14.187
(D) N-BWE I	**0.443**
(E) N-BWE II	0.643

B. Experimental results

To compare the algorithmic latencies of each BWE method, the latencies were calculated under the 16-kHz-sampled scenario ($K = 2, F_{s_0} = 8$ kHz, $F_{s_1} = 16$ kHz) as shown in Table I. From the results, (D) N-BWE I achieved the lowest latency. This was because (D) used only the band-pass filter with small number of filter orders. The latencies of (B) and (E) were almost similar to (D) because the filter order strongly depended on the latency. (B) and (E) still satisfied the requirements of real-time applications. In the case of (C) LPAS, the number of FFT points was also included in the calculation. The LPAS algorithm was implemented with a 2048-point FFT, and the latency of FFT only was 12.8 ms.

Figure 4 illustrate the results of each objective measurement with box plots [6] and shows the perceptual evaluation of speech quality (PESQ) [11], a short-time objective intelligibility (STOI) [12], and RMS-LSD scores for each BWE method. From the results, N-BWE obtained a slightly higher PESQ



(a) PESQ (b) STOI

(c) RMS-LSD

Fig. 4: Objective measurements

scores, and a lower RMS-LSD value than LPAS. STOI was one the measurement for the naturalness. It can be seen that there are some relationships between STOI scores and their latency. From these results, it was indicated that the N-BWE method was able to lead to better RMS-LSD values with low-latency. Consequently, it turned out that the N-BWE method was effective for real-time applications.

V. Conclusions

This paper evaluated the effects of some non-learning and blind BWE methods. The N-BWE is a blind, non-learning and lightweight BWE approach. For real-time applications such as video calls, low-latency BWE methods that are applicable to several network services are strongly required. From the experimental results, the N-BWE method provided the lowest latency and better RMS-LSD values among the conventional BWE methods. In future work, the BWE methods can use as a technique for data augmentation, the effectiveness will be evaluated.

Acknowledgment

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

References

[1] 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.

[2] 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.

[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, no. 25, pp. 2149–2151, 2015.

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

[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. 5429–5433, 2018.

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

[7] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Blind bandwidth extension with a non-linear function and its evaluation on x-vector-based speaker verification (acceoted)," *in Proc. INTERSPEECH*, 2019.

[8] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Jounal pf the American Academiy Audiology 11*, pp. 330–336, 2000.

[9] D. Zaykovskiy and B. Iser, "Comparison of neural networks and linear mapping in an application for bandwidth extension," *in Proc. of SPECOM*, pp. 1–4, 2005.

[10] M. Mitchell, F. Luciana, C. Diego, and L. Aaron, "The speakers in the wild (sitw) speaker recognition database," *in Proc. INTERSPEECH*, pp. 818–822, 2016.

[11] 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.

[12] 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.