

Artificial Bandwidth Extension for Speech Signals using Speech Recognition

Shingo Kuroiwa¹, Masashi Takashina¹, Satoru Tsuge¹, Ren Fuji^{1,2}

¹ Department of Information Science and Intelligent Systems, The University of Tokushima, Tokushima 770-8506, Japan

² School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

{kuroiwa, tsuge, ren}@is.tokushima-u.ac.jp

Abstract

In this paper, we propose a non-realtime speech bandwidth extension method using HMM-based speech recognition and HMM-based speech synthesis. In the proposed method, first, the phoneme-state sequence is estimated from the bandlimited speech signals using the speech recognition technique. Next, for estimating spectrum envelopes of lost high-frequency components, an HMM-based speech synthesis technique generates a synthetic speech signal (spectrum sequence) according to the predicted phoneme-state sequence. Since both speech recognition and speech synthesis take into account dynamic feature vectors, we can obtain a smoothly varying spectrum sequence. For evaluating the proposed method, we conducted subjective and objective experiments. The experimental results show the effectiveness of the proposed method for bandwidth extension. However, the proposed method needs more improvement in speech quality.

Index Terms: bandwidth extension, speech recognition, HMM, dynamic features

1. Introduction

In general, bandwidth limited speech signals are hard to catch for humans. For example, when speech is transmitted through telephone lines, both the high and low frequency speech components are lost by limiting the frequency bandwidth. Therefore, speech clarity and speech presence are lost, and deterioration of speech quality takes place. Low quality speech caused by narrowed frequency bandwidth is called bandlimited speech. It is important to improve the speech quality of a bandlimited speech for communications.

In recent years, various methods have been proposed for improving degradation of speech quality caused by bandlimiting[1]. In bandwidth extension (BWE) methods, the lost components are artificially estimated and generated from the bandlimited speech information and wideband speech is generated by complementing the estimated components to the bandlimited speech. A typical BWE method uses the technique of linear mapping[2], codebook mapping[3] and layered neural network[4]. Statistical methods, such as GMM[5] and HMM[6], are also applied for accurate estimation in BWE. However, while the generated wideband speech of the conventional methods improves the feeling of the bandwidth extension, it is necessary to improve the speech quality. To improve the generated wideband speech quality, it is necessary to generate the lost components more accurately.

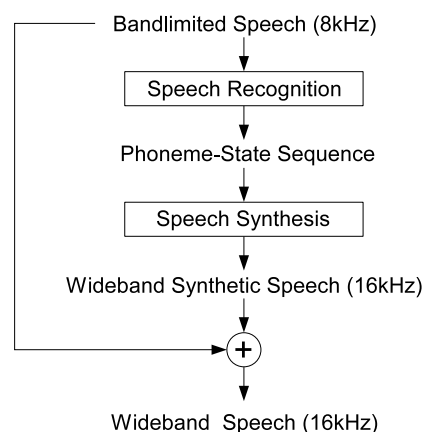


Figure 1: Schematic diagram of the proposed BWE method

In this paper, we propose a non-realtime BWE method that utilizes HMM-based speech recognition and HMM-based speech synthesis. In the method, first, wideband synthetic speech is generated by speech recognition and speech synthesis using bandlimited speech information. Next, the lost components are extracted from the generated synthetic speech, and then the extended bandwidth speech is generated by complementing the extracted components with the bandlimited speech. By using HMM-based speech recognition and speech synthesis, we can generate wideband synthetic speech with high accuracy from bandlimited speech. Since both speech recognition and speech synthesis take into account dynamic feature vectors, we can obtain a smoothly varying spectrum sequence. Hence, by using synthetic speech information as the lost components, it is possible to generate high quality and more natural wideband speech. For evaluating the validity of the proposed methods, both objective and subjective experiments that looked at the quality of speech and the improvement of feeling in bandwidth extension were conducted.

2. BWE using Speech Recognition

2.1. Outline of the proposed method

A block diagram of the proposed method is shown in Figure 1. First, an HMM-based speech recognizer with a speaker-independent HMM and an N-gram language model predicts the phoneme-state sequence from the inputted bandlimited speech signal. Next, an HMM-based speech synthesizer[7] with a

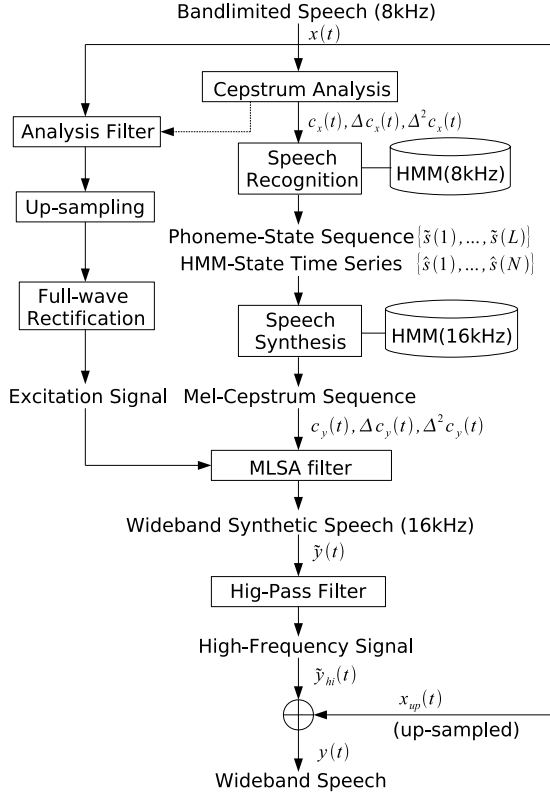


Figure 2: speech bandwidth extension method using synthetic speech waveform

speaker-independent HMM generates a spectrum envelope sequence according to the predicted phoneme-state sequence. The excitation signal is generated by full-wave rectification of the bandlimited speech source signal. Next, using this information (spectrum envelope sequence and speech source signal data), the MLSA (Mel Log Spectral Approximation) filter[8] synthesizes the wideband synthetic speech. Finally, the wideband speech is generated by adding the inputted bandlimited speech and high-frequency components of the synthetic speech.

2.2. Algorithm

Here, we will give the full details of the algorithm. Figure 2 shows a more precise diagram of the proposed method.

Let $\mathbf{X}$ and $\mathbf{Y}$ be a bandlimited speech frame sequence and a corresponding wideband speech frame sequence, respectively.

$$\mathbf{X} = \{\mathbf{x}(1), \dots, \mathbf{x}(t), \dots, \mathbf{x}(N)\}, \quad (1)$$

$$\mathbf{Y} = \{\mathbf{y}(1), \dots, \mathbf{y}(t), \dots, \mathbf{y}(N)\}, \quad (2)$$

where, $\mathbf{x}(t)$ and $\mathbf{y}(t)$ is the framed speech at time t and N is the number of frames. The objective of BWE can be formulated as estimating $\tilde{\mathbf{Y}}$ so that it maximizes the following joint probability.

$$\tilde{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} P(\mathbf{Y}, \mathbf{X}), \quad (3)$$

$$= \underset{\mathbf{Y}}{\operatorname{argmax}} P(\mathbf{Y} | \mathbf{X})P(\mathbf{X}), \quad (4)$$

Because $P(\mathbf{X})$ is common regardless of $\mathbf{Y}$, the equation can be simplified to

$$\tilde{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} P(\mathbf{Y} | \mathbf{X}). \quad (5)$$

Since it is very difficult to find the $\tilde{\mathbf{Y}}$ directly, we propose to use speech recognition and speech synthesis. $\tilde{\mathbf{Y}}$ is estimated by the following steps:

- (1) the mel-frequency cepstral coefficients at frame t , $\mathbf{c}_x(t)$ are derived from the inputted bandlimited speech frame. The dynamic features $\Delta\mathbf{c}_x(t)$ and $\Delta^2\mathbf{c}_x(t)$, i.e. delta and delta-delta mel-cepstral coefficients at frame t , are also calculated. Then, the feature vector of each frame, $\mathbf{C}_x(t)$, is represented respectively as follows:

$$\mathbf{C}_x(t) = [\mathbf{c}_x(t)^\top, \Delta\mathbf{c}_x(t)^\top, \Delta^2\mathbf{c}_x(t)^\top]^\top, \quad (6)$$

where, $\top$ denotes transposition of the vector. The excitation signal is also extracted using an analysis filter based on the mel-frequency cepstrum.

- (2) The HMM-based speech recognizer with a band-limited speaker-independent HMM and an N-gram language model finds the phoneme-state sequence (HMM state sequence), $\{\tilde{s}(1), \dots, \tilde{s}(L)\}$ which maximizes the following conditional probability:

$$P(\{\tilde{s}(1), \dots, \tilde{s}(L)\} | \{\mathbf{C}_x(1), \dots, \mathbf{C}_x(N)\}), \quad (7)$$

where, L is the number of phoneme-states and is smaller than N .

- (3) The Viterbi algorithm, employed by the speech recognizer estimates the HMM-state corresponding to the observed speech frame sequence $\mathbf{X}$. This calculation is simultaneously carried out during the speech recognition process. As a result, the predicted HMM-state time series,

$$\hat{\mathbf{S}} = \{\hat{s}(1), \dots, \hat{s}(N)\}, \quad (8)$$

is obtained.

- (4) The HMM-based speech synthesizer[7] with a wide-band speaker-independent HMM generates a cepstrum sequence, $\{\mathbf{C}_y(1), \dots, \mathbf{C}_y(N)\}$, that maximizes the following conditional probability:

$$P(\{\mathbf{C}_y(1), \dots, \mathbf{C}_y(t), \dots, \mathbf{C}_y(N)\} | \hat{\mathbf{S}}), \quad (9)$$

where,

$$\mathbf{C}_y(t) = [\mathbf{c}_y(t)^\top, \Delta\mathbf{c}_y(t)^\top, \Delta^2\mathbf{c}_y(t)^\top]^\top. \quad (10)$$

Without dynamic features, the cepstrum sequence that maximizes equation (9) is a sequence of the mean vectors of corresponding HMM-states. However, a smoothly varying cepstrum sequence, or a spectrum envelope sequence, is obtained by taking into account dynamic features [9]. Finally, using the estimated cepstrum sequence, the MLSA filter[8] synthesizes the wideband synthetic speech, $\tilde{\mathbf{y}}(t)$, using the excitation signal obtained in step (1). For extension of the excitation signal, up-sampling and full-wave rectification are applied.

- (5) Using a high-pass filter, $H_{hp}(\omega)$ (3.4kHz - 8kHz), a high-frequency waveform, $\tilde{\mathbf{y}}_{hi}(t)$, is extracted from the synthetic speech, $\tilde{\mathbf{y}}(t)$. Finally, the wideband speech signal, $\mathbf{y}(t)$, is obtained by adding the extracted waveform, $\tilde{\mathbf{y}}_{hi}(t)$ to the up-sampled input speech signal, $\mathbf{x}_{up}(t)$.

$$\mathbf{y}(t) = \mathbf{x}_{up}(t) + \tilde{\mathbf{y}}(t). \quad (11)$$

Table 1: Acoustic analysis conditions for speech recognition

sampling frequency	8kHz
frame length	25ms
frame shift	10ms
window style	Hamming
feature vector	1-12 mel-cepstral coeffs.(CMS), Δ el-cepstral coeffs., Δ LogPower(total 25)

Table 2: Acoustic analysis conditions for speech synthesis

sampling frequency	16kHz
frame length	25ms
frame shift	5ms
window style	Blackman
frequency warping factor	0.42
feature vector	0-18 mel-cepstral coeffs., Δ mel-cepstral coeffs., Δ^2 mel-cepstral coeffs.,(total 57)

3. Experiments

To evaluate the proposed method, objective and subjective experiments were conducted. For comparison, a BWE method based on codebook mapping was also evaluated.

3.1. Experimental conditions

3.1.1. Speech recognition

Training data consisted of 2,700 utterances from 54 female speakers in the JNAS database[10]. These training data were preliminarily down-sampled by using sox[14]. The acoustic analysis conditions are shown in Table 1. For acoustic models, shared-state triphone HMMs with 16 Gaussian mixture components per state were trained using HTK[15]. In total, there were 1,100 HMM-states. We used Julius[11, 16] for the speech recognizer with a 20,000 word lexicon and a *tri*-gram language model.

3.1.2. Speech synthesis

The speaker-independent HMM for speech synthesis was trained with the same 2,700 utterances from 54 female speakers. The acoustic analysis conditions are shown in Table 2, which were different from the ones used for recognition. For acoustic models, shared state triphone HMMs with a single Gaussian per state were trained. As with speech recognition there was a total of 1,100 HMM-states. We used SPTK[17] for acoustic analysis and speech synthesis, and HTK[15] for training the acoustic model.

3.1.3. Codebook mapping

For comparison, a BWE method based on codebook mapping [3] was implemented. Table 3 shows experimental conditions used with this method.

3.1.4. Test set

For the test set, we used ten utterances from a female speaker who was not included among the HMM training speakers. For

Table 3: Experimental conditions for codebook mapping

frame length	25ms (16kHz) 12.5ms (8kHz)
frame shift	5ms (16kHz) 2.5ms (8kHz)
window style	Hamming
LPC analysis order	24 (16kHz) 12 (8kHz)
distance measure	LPC cepstrum distance
codebook size	256

Table 4: Evaluation standard used in CMOS

subjective standard	opinion score
The former is much better than the latter	2
The former is better than the latter	1
About the same	0
The latter is better than the former	-1
The latter is much better than the former	-2

simulating the bandwidth limitation, bandlimited speech (0 - 3.4kHz) was artificially generated in each utterance by down-sampling (16kHz to 8kHz) using sox[14].

3.2. Objective evaluation

For the objective evaluation, the segmental SNR (SNRseg) was measured using the original wideband speech and the created bandwidth extended speech to compare the speech quality between the two. SNRseg is calculated using the following equation.

$$\text{SNR}_{\text{seg}} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left(\frac{\sum_{n=0}^{L-1} x(j, n)^2}{\sum_{n=0}^{L-1} (\tilde{x}(j, n) - x(j, n))^2} \right), \quad (12)$$

where, $x(j, n)$ and $\tilde{x}(j, n)$ is the j -th original speech frame (16kHz) and the bandwidth extended speech at time n respectively, L is the number of sample data for each frame, and M is the number of frames.

3.3. Subjective evaluation

Subjective evaluation was expressed in five stages and a comparison mean opinion score (CMOS)[12] was calculated. In the subjective experiment, the generated wideband speech and the bandlimited speech were presented to eight listeners in order of “the bandlimited speech → the wideband speech” or “the wideband speech → the bandlimited speech” randomly. The listeners assigned scores for “feeling of bandwidth extension” and “speech quality” based on Table 4. Before the experiments, an example of “feeling of bandwidth extension” was presented to the listeners using a bandlimited speech example with its original wideband version.

3.4. Experimental results

Table 5 shows the objective evaluation results. These values are mean values of SNRseg for each method. From the table, we can see that the proposed method is able to improve the speech quality over the conventional codebook mapping algorithm a little.

Table.6 shows the subjective evaluation results. These values are the mean CMOS values for each method. From the

Table 5: Results of the objective evaluation experiments (distortion to the original speech (SNRseg))

method	SNR _{seg}
proposed method	17.2 (dB)
codebook mapping	15.6 (dB)

Table 6: Results of the subjective evaluation experiments (average CMOS)

method	speech quality	speech bandwidth
proposed method	-0.5	+0.7
codebook mapping	-1.0	+0.9

table, we can see that both methods were able to improve the bandwidth extension. The proposed method had slightly better speech quality, but the result is still not as good as we hoped. The listeners found the proposed method created noise in the speech.

Figure 3 shows the spectrogram of speech in the proposed method. The spectrum envelopes of the proposed method seem to be more flat than the original ones. This result is probably brought by the use of the speaker-independent single Gaussian HMM for the speech synthesis. Hence, we will attempt to use the speaker-adapted multi-mixture HMM in the future. We will also consider a spectral conversion technique [13] that does not consider phoneme as an intermediate for effective bandwidth extension. We also have to improve the extension method of the excitation signal, which is the probable cause of noisy impression the listeners had.

4. Conclusion

We proposed a non-realtime speech bandwidth extension method using speech recognition and speech synthesis. In the proposed method, the spectrum envelope sequence was estimated by means of maximum likelihood taking into account dynamic features. In the objective experiments, the proposed method showed an improvement of 1.6 dB in the segmental SNR compared with conventional codebook mapping. In the subjective experiment, the proposed method showed similar “feeling of bandwidth extension” and a little improvement in “speech quality” compared with codebook mapping. However, the speech quality is still not good enough, because of listeners impression of noise that is probably caused by the excitation signal. Hence, we will attempt to improve the extension method of the excitation signal.

5. Acknowledgment

This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 17300036, 17300065, Exploratory Research, 17656128, 2007.

6. References

- [1] E. Larsen, R.M. Aarts, Audio Bandwidth Extension, John Wiley & Sons Ltd., Chichester, 2004.
- [2] Y. Nakatoh, M. Tsushima, T. Norimatsu, “Generation of broadband speech from narrowband speech using piecewise linear mapping”, *EUROSPEECH*, pp.1643–1646, 1997.
- [3] Y. Yoshida, M. Abe, “An algorithm to reconstruct wideband speech

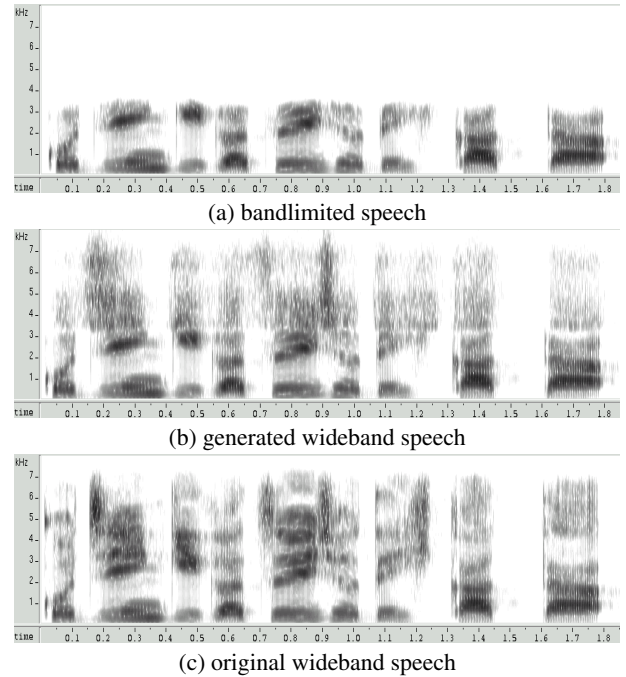


Figure 3: spectrogram on the proposed method

- from narrowband speech based on codebook mapping”, *ICSLP*, pp.1591–1594, 1994.
- [4] Y. Tanaka, N. Hatazoe, “Reconstruction of wideband speech from telephone-band speech by multi-layer Neural Networks”, *ASJ meeting*, 1-4-19, pp.255–256, 1995. (in Japanese)
- [5] K.Y. Park, H.S. Kim, “Narrowband to wideband conversion of speech using GMM-based transformation”, *ICASSP*, pp.1847–1850.
- [6] P. Jax, P. Vary, “On artificial bandwidth extension of telephone speech”, *Signal Processing*, Vol.83, pp.1707–1719, 2003.
- [7] K. Tokuda, T. Yoshimura, T. Masauko, T. Kobayashi, “Speech parameter generation algorithms for HMM-based speech synthesis”, *Proc. IEEE*, vol.3, pp.1315–1358, 2000.
- [8] S. Imai, “Cepstral analysis synthesis on the mel frequency scale”, *ICASSP*, pp.93–96, 1983.
- [9] T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, “Speech synthesis using HMMs with dynamic features”, *ICASSP*, pp.389–392, 1996.
- [10] K. Itou, M. Yamamoto, K. Takeda, T. akezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, “The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus”, *ICSLP*, pp.3261–3264, 1998.
- [11] A. Lee, T. Kawahara and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine”, *EUROSPEECH*, pp.1691–1694, 2001.
- [12] Keagy, S., *Intrgrating Voice and Data Networks*. Cisco Press, 2000.
- [13] T. Toda, A. Black, K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter”, *ICASSP*, pp.9–12, 2005.
- [14] <http://sox.sourceforge.net/>
- [15] <http://htk.eng.cam.ac.uk/>
- [16] <http://julius.sourceforge.jp/>
- [17] <http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/>