View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Crossref

Chen *et al. EURASIP Journal on Audio, Speech, and Music Processing* 2013, **2013**:10
http://asmp.eurasipjournals.com/content/2013/1/10

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An audio watermark-based speech bandwidth extension method

Zhe Chen, Chengyong Zhao, Guosheng Geng and Fuliang Yin*

Abstract

A novel speech bandwidth extension method based on audio watermark is presented in this paper. The time-domain and frequency-domain envelope parameters are extracted from the high-frequency components of speech signal, and then these parameters are embedded in the corresponding narrowband speech bit stream by the modified least significant bit watermark method which uses perception property. At the decoder, the wideband speech is reproduced with the reconstruction of high-frequency components based on the parameters extracted from bit stream of the narrowband speech. The proposed method can decrease poor auditory effect caused by large local distortion. The simulation results show that the synthesized wideband speech has low spectral distortion and its speech perception quality is greatly improved.

1 Introduction

The narrowband speech with 8 KHz sampling frequency is widely used in many communication systems [1]. This kind of speech sounds unnatural due to the missing of high-frequency components; therefore, it can not meet the demands for high-quality perception, such as telephone/video conference systems. With the increasing of communication network bandwidth, wideband speech transmission is strongly desired, but large-scale update of narrow communication infrastructures is difficult and expensive. For the existing communication network, such as public switched telephone network (PSTN) and global system for mobile communication (GSM), speech bandwidth extension (BWE) technique is an effective and realistic choice to obtain wideband speech quality.

Speech BWE methods are mainly divided into two classes. One is based on correlation between narrowband speech components and wideband ones; the other is based on information hiding technique. Most of the former methods produce wideband speech by linear prediction (LP) model [2], i.e., excitation signal and linear prediction coefficients (stand for spectral envelope). Nagel et al. proposed high-frequency (HF) information generation method based on signal sideband modulation [3], i.e., low-frequency (LF) band signal is first modulated, then extended into HF part, and, finally, filled the gap between LF and HF with noise and shaped the frequency-domain envelope. Fuchs and Lefebvre proposed a harmonic BWE method [4]. This method generated HF components by parallel phase vocoder and removed noise in the intersection part of spectrums. Pulakka et al. proposed a speech BWE method using Gaussian mixture model based estimation of the high band Mel spectrum [5]. Pulakka and Alku proposed a BWE method of telephone speech using neural network and filter bank implementation for high-band Mel spectrum [6]. Pham et al. used back-forward filter to generate excitation signal [7], which makes perception quality of synthesized wideband speech improve greatly. Bauer and Fingscheidt used pre-trained neural network to generate HF speech components and synthesized wideband speech by spline interpolation method [8]. Naofumi proposed a hidden Markov model (HMM)-based BWE methods [9]. This method can enhance the speech quality without increasing the amount of transmission data. These methods, based on correlation between narrowband speech components and wideband ones, have low enough computational complexity, but noises are easily introduced into the frequency band between LF and HF [10].

The speech BWE methods based on information hiding technique usually embed HF components information into the bit stream of narrowband speech, and then, the wideband speech is recovered based on the HF information at the receiver. Chen and Leung proposed a

*Correspondence: flyin@dlut.edu.cn
School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China

Springer
© 2013 Chen et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

speech BWE method based on least significant bits (LSB) audio watermark [11], which can embed more HF speech components information but is susceptible to noise and channel interference. Geiser and Vary proposed a speech BWE method based on data hiding technique [12]. They embedded linear prediction coefficients of HF components into the encoded narrowband speech then recovered the data in the decoder and synthesized wideband speech. But when suffering from the channel interference, this method has poor synthesized wideband speech. Esteban and Galand proposed a speech BWE method based on the GSM EFR codec [13], which embed the sideband information into the narrowband speech stream by watermark. This method can synthesize wideband speech with less noise.

In this paper, a new BWE method based on the modified LSB watermark technique is proposed. This method first extracts the necessary HF components parameters, including time-domain envelopes, frequency-domain envelopes, and energy of the wideband speech; then these parameters are compressed and embedded into the narrowband speech bit stream with a modified watermark technique. In decoder, the reverse procedure is applied to extract the HF parameters; then these parameters are used to synthesize HF components; finally, the wideband speeches are recovered from the LF and HF speech components.

2 Speech BWE method based on audio watermark

The block diagram of the proposed BWE method is shown in Figure 1, including quadrature mirror filter (QMF) based analysis filter bank, down-sampler, HF parameters extractor, G.711 encoder, watermark embedder at transmitting terminal, G.711 decoder, watermark extractor, HF speech restorer, up-sampler, and QMF synthesis filter bank. At the receiving terminal, from Figure 1, first, input wideband speech with 16-KHz sampling frequency is put into two-channel QMF bank [14], and filter bank's outputs are down-sampled twice. Thus both HF and LF components with 8-KHz sampling frequency are obtained. Second, the LF components are encoded by the G.711 encoder. The HF parameters are estimated from the HF components by HF parameters extractor. Third, HF parameters are compressed and embedded into G.711 bit stream by modified watermark method, and the bit stream-embedded HF parameters are transmitted to the receiver through a narrowband communication network. At the receiving terminal, narrowband speech is decoded with G.711 decoder, while the HF parameters are extracted from the received bit stream, and then the HF speech is recovered with HF parameters. After recovering both LF and HF speech components, their sampling frequency is doubled, and the wideband speech is finally synthesized through two-channel QMF filter-based synthesis bank. Every module in Figure 1 will be discussed in detail in the following subsections.

2.1 Down-sampling processing of speech signal

Here the analysis filter bank used in Recommendation G.729.1 is adopted [14]. There are two filters in the filter bank, i.e., low-pass filter (LPF) and high-pass filter (HPF). Their unit impulse responses are $h_{\mathrm{L}}(n)$ and $h_{\mathrm{H}}(n)$ respectively. LPF's technical specifications can be summarized as (a) sampling frequency, 16 KHz; (b) passband cutoff frequency, 3.7 KHz; (c) stopband cutoff frequency, 4.5 KHz; (d) maximum passband ripple, 0.015 dB; and (e) the minimum stopband attenuation, 39 dB. According to QMF filter bank theory, the unit impulse responses of HPF is $h_{\mathrm{H}}(n) = h_{\mathrm{L}}(n)\mathrm{e}^{jn\pi} = (-1)^n h_{\mathrm{L}}(n)$. The frequency responses of LPF and HPF are dot-solid line and solid line in Figure 2, respectively.

The QMF analysis filter bank divides the wideband speech into two parts: 0 to 4 KHz LF components and 4 to 8 KHz HF components. To remove redundant information, the sampling frequency of both LF and HP components is reduced to 8 KHz by down-sampler. Thus, the LF components $s_{\mathrm{L}}(n)$ and HF components $s_{\mathrm{H}}(n)$ can be expressed as

$$s_{\mathrm{L}}(n) = \sum_{m=0}^{\mathrm{ORD}-1} s_{\mathrm{wb}}(2n-m)h_{\mathrm{L}}(m), \qquad n = 0, 1, \ldots \tag{1}$$

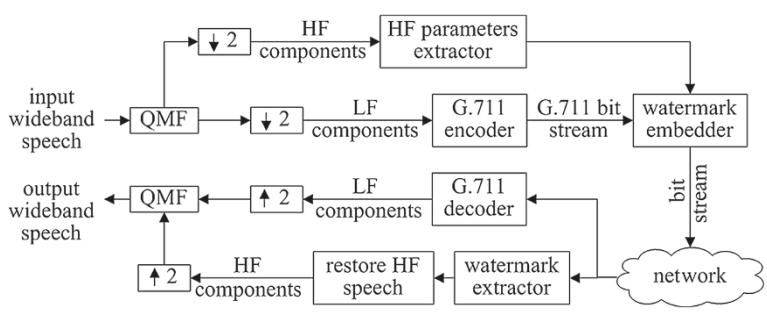

Figure 1 Block diagram of proposed speech BWE scheme.

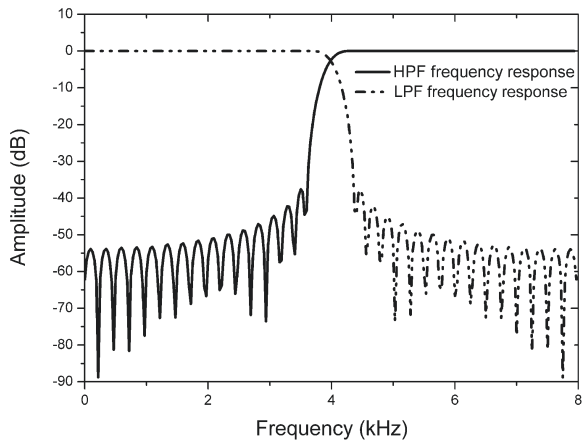

Figure 2 Amplitude-frequency responses of LPF and HPF.

$$s_{\mathrm{H}}(n) = \sum_{m=0}^{\mathrm{ORD}-1} s_{\mathrm{wb}}(2n-m)h_{\mathrm{H}}(m), \qquad n = 0, 1, \ldots, \tag{2}$$

where the filter order ORD is equal to 64, and s_{wb} is the input wideband speech signal.

2.2 High-frequency parameters extraction

The parameters of HF components include the time-domain and frequency-domain envelopes and their averages. First, a HF speech frame, including 160 samples, is divided into 16 segments, i.e., each segment has 10 samples. The time-domain envelope of the ith segment $T(i)$ can be calculated as [14]

$$T(i) = \frac{1}{2}\log_2[\sum_{n=0}^{9} s_{\mathrm{H}}^2(n+10i)], \qquad i = 0, 1, \ldots, 15. \tag{3}$$

The average M_{T} of $T(i)$ can be obtained [14]

$$M_{\mathrm{T}} = \frac{1}{16}\sum_{i=0}^{15} T(i). \tag{4}$$

To remove M_{T} from $T(i)$ [15], the time-domain envelope $T_{\mathrm{M}}(i)$ is

$$T_{\mathrm{M}}(i) = T(i) - M_{\mathrm{T}}, \qquad i = 0, 1, \ldots, 15. \tag{5}$$

By applying semi-Hamming window to a HF speech components and then attaching zero samples until the total samples number reaches 256 [14], we have

$$\begin{cases} S_{\mathrm{H}}^{w}(n) = w(n)S_{\mathrm{H}}(n), & n = 0, \ldots, 159 \\ S_{\mathrm{H}}^{w}(n) = 0, & n = 160, \ldots, 255, \end{cases} \tag{6}$$

where semi-Hamming window $w(n)$ is

$$w(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n/96), & n = 0, \ldots, 47 \\ 1, & n = 48, \ldots, 159. \end{cases} \tag{7}$$

After fast Fourier transform (FFT), we have

$$\begin{aligned} S_{\mathrm{H}}(k) &= \mathrm{FFT}[s_{\mathrm{H}}^{w}(n)] \\ &= \sum_{n=0}^{L-1} s_{\mathrm{H}}^{w}(n)\mathrm{e}^{-j\frac{2\pi}{L}kn}, \qquad k = 0, 1, \ldots, L-1, \end{aligned} \tag{8}$$

where $L = 256$.

The frequency band of HF speech is uniformly divided into 12 intervals. In order to reduce the range of parameters and take the difference of the contribution of each point in the interval into account, the 12 frequency bands information are converted to weighted energy in sub-band, also named frequency envelope. The frequency envelope $F(k)$ for the kth interval is calculated as [14]

$$F(k) = \frac{1}{2}\log_2[\sum_{i=10k}^{10k+11} w_{\mathrm{H}}(i-2k)|S_{\mathrm{H}}(i)|^2], \quad k = 0, 1, \ldots, 11 \tag{9}$$

where the weighting window w_{H} of sub-band frequency domain is defined as

$$w_{\mathrm{H}}(n) = \begin{cases} 1, & n = 1, 2, \ldots, 10 \\ 0.5, & n = 0, 11. \end{cases} \tag{10}$$

The average frequency-domain envelope M_{F} is

$$M_{\mathrm{F}} = \frac{1}{12}\sum_{k=0}^{11} F(k). \tag{11}$$

Subtracting M_{F} from $F(k)$, the frequency-domain envelope $F_{\mathrm{M}}(k)$ is obtained as [15]

$$F_{\mathrm{M}}(k) = F(k) - M_{\mathrm{F}}, \qquad k = 0, 1, \ldots, 11. \tag{12}$$

2.3 Watermark embedding and extracting

In each speech frame, the number of HF parameters is 30, including 16 time-domain envelope ($T_{\mathrm{M}}(i)$, $i = 0, 1, \ldots, 15$), 12 frequency-domain envelope ($F_{\mathrm{M}}(k)$, $k = 0, 1, \ldots, 11$), average time-domain envelope M_{T}, and average frequency-domain envelope M_{F}. Usually, these raw M_{T} and M_{F} are floating-point format, whereas embedded watermark is regarded as binary numbers, so the floating-point numbers need to be converted to binary ones. To reduce the deviation by bits error, conversion precision is set to 12 bits, where the former 6 bits represent the integer part, the latter 6 bits represent the fractional part multiplied by 32. A typical representation of the watermark data is shown in Figure 3.

In order to further reduce the amount of data, vector quantization (VQ) is conducted to both time-domain

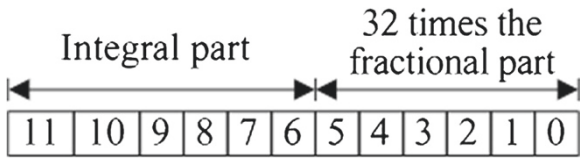

Figure 3 The average time-domain and frequency-domain envelope in watermark data.

and frequency-domain envelopes [16]. In the VQ process, the time-domain and frequency-domain envelopes are divided into four sections and three sections, respectively, where each section is a four-dimensional vector and is quantized with 6 bits. Thus, the total number of digital information is $12 + 12 + 6 * 4 + 6 * 3 = 66$ bits, and the quantization code book in reference [14] is available.

Usually, audio watermark is designed to be undetectable and perceivable but can be extracted with a hidden message by some algorithms. Using this feature of watermark, we assign the 66 bits digital information as watermark and embed it into LF bit stream; thus in the receiving terminal, HF information hidden can be obtained with watermark extractor. In this paper, a modified LSB watermark method is proposed, which is based on communication protocol characteristics and human hearing perception.

According to the time-domain masking effect of human auditory, a large signal can make masking effect on the small signal [1]. So changes in the small signals can not be easily heard. With this auditory characteristics, we embed the watermark with LF and HF components parameters into the small signal position to make the watermark hidden better.

The detailed modified watermark method is as follows: C0 to C7 indicate the encoded bit stream from the lowest to the highest position, as shown in Figure 4. According to G.711 codec format, C7 is the symbol bit of the sampling points. We uses C6 to distinguish large-signal (C6 = 1) with small signal (C6 = 0), thus when C6 is equal to 0, the watermark is embedded. If embedded position is less than 66 bits, the other positions must be chosen to embed watermark.

When extracting watermark, we decide whether watermark is embedded or not based on the characteristics of bit streams. If the C6 bit is 0, the watermark is extracted from the lowest position of bits; if the C6 bit is 1, there is no watermark in bit stream. If reaching the end of the frame but the extracted watermarks are less than 66 bits, then return to a starting point and extract watermark in the C6 = 1 position until the watermark bits extracted are up to 66 bits.

2.4 Recovery of HF components

The block diagram of HF components recovery is shown in Figure 5. Because the HF components and LF ones have correlation more or less [17], the LF components is used to construct the autoregressive (AR) model with transfer function $H(z)$ [18]

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{13}$$

where a_i is linear prediction coefficient of the LF part, p is the order of AR model, G is the gain.

In the decoder, white noise signal is generated as [18]

$$\text{seed}(n) = (\text{word16})\,[31,821 \cdot \text{seed}(n-1) + 13,849] \tag{14}$$

where (word16) is the operation reserving lower 16 bits only, and the random seed, seed(n), at n time is a 16-bit integer and its initial value is 12,357. Let seed(n) through the AR model given in Equation 13, i.e.,

$$u(n) = G\,\text{seed}(n) + \sum_{i=1}^{p} a_i u(n-i). \tag{15}$$

When obtaining $u(n)$ from the AR model, the parameters of HF components are also extracted from watermark in LF bit stream, including 16 time-domain envelopes, 12 envelope frequency-domain envelopes, the average time-domain envelope, and the average frequency-domain envelope. Then, the HF parameters recovered from LF bitstream are used to shape both time-domain and frequency-domain envelopes of $u(n)$ [15]. Since shaping method of the frequency-domain envelope is similar with the one in time-domain, shaping process of time-domain envelope is only given as follows.

From the extracted watermark, we can build the time-domain envelope $T_{\mathrm{M}}(i)$ and the average time-domain envelope M_{T}. Then time-domain envelope of HF components are recovered as

$$T(i) = T_{\mathrm{M}}(i) + M_{\mathrm{T}}, \qquad i = 0, 1, \ldots, 15 \tag{16}$$

The local gain factors of time-domain are computed as

$$\text{gain_}t(i) = 2^{T(i) - \tilde{T}(i)}, \qquad i = 0, 1, \ldots, 15, \tag{17}$$

where $\tilde{T}(i)$ are the envelope parameters of $u(n)$ in time domain.

The gain factor between the two fragments can be obtained with linear interpolation

$$\text{gain}(n + 10i) = \begin{cases} \frac{1}{9}[\,\text{gain_}t(i) - \text{gain_}t(i-1)]\,(n-4) + \text{gain_}t(i), & n = 0,\ 1,\ 2,\ 3 \\ \text{gain_}t(i), & n = 4,\ 5 \\ \frac{1}{9}[\,\text{gain_}t(i+1) - \text{gain_}t(i)]\,(n-5) + \text{gain_}t(i), & n = 6,\ 7,\ 8,\ 9. \end{cases} \tag{18}$$

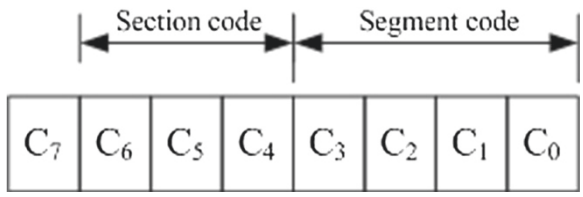

Figure 4 G.711 bit stream format.

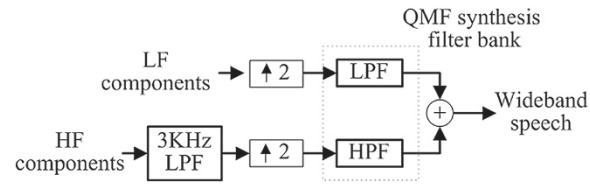

Figure 6 Block diagram of wideband speech synthesis.

The time-domain envelope of noise $u(n)$ can be adjusted by local gain factor

$$u_t(n+10i) = u(n+10i)\,\mathrm{gain}(n+10i), \quad n = 0, 1, \ldots, 9 \quad i = 0, 1, \ldots, 15. \tag{19}$$

After above-mentioned time-domain and frequency-domain envelopes are shaped, the HF speech components are reconstructed.

2.5 Synthesis of wideband speech

The block diagram of wideband speech synthesis is shown in Figure 6. With G.711 decoder, the receiving bit stream is decoded to LF components with sampling frequency of 8 KHz. In order to remove the uncomfortable noise above 7 KHz, the reconstructed HF components are filtered with a low-pass filter, whose technical specifications can be summarized as (a) passband cutoff frequency, 3 KHz; (b) stopband cutoff frequency, 3.4 KHz; (c) maximum passband ripple, 0.8 dB; (d) minimum stopband attenuation, 80 dB. The LF components and filtered HF components are up-sampled to 16 KHz by twice interpolation and then are synthesized to a wideband speech with QMF synthesis filter bank, which is the reciprocal of QMF analysis filter bank in Section 2.1.

3 Simulation and result discussion

In order to evaluate the performance of proposed BWE scheme, both objective and subjective experiments are carried out. Without loss of generality, according to the character of pitch and timbre, test speeches are divided into five types: male speech, female speech, boy speech, girl speech, and song. All test speeches are quantized with 16 bits and sampled at 16 KHz. These speeches will be used as the original wideband speeches for the following experiments.

3.1 Objective measurements

The objective measurements, including spectral distortion and spectrogram, are used to compare the performance between original wideband speech at transmitting terminal and expanded wideband speech at receiving terminal.

The spectral distortion D_{HC} is defined as [19]

$$D_{\mathrm{HC}}^2 = \frac{1}{K}\sum_{k=1}^{K}\int_{0.5\pi}^{\pi}\left[20\lg\left(\frac{A_k(\omega)}{A'_k(\omega)}\right) - G_{\mathrm{C}}\right]^2 d\omega, \tag{20}$$

where $A_k(\omega)$ and $A_k^{'}(\omega)$ are the kth frame spectral envelopes for the original wideband speech and expanded wideband speech respectively, G_{C} is the gain compensation factor for removing the mean squared error between the two envelopes and is defined as

$$G_{\mathrm{C}} = \frac{1}{0.5\pi}\int_{0.5\pi}^{\pi} 20\lg\left(\frac{A'_k(\omega)}{A_k(\omega)}\right)d\omega. \tag{21}$$

We select the five types of speech mentioned above with 52 s length and calculate their spectral distortion. Experience results of spectral distortion are shown in Table 1. Usually, the smaller the spectral distortion is, the more similar the synthesis of wideband speech and original speech is. From Table 1, an interesting result can be found that the spectral distortion of song is lower than the speech.

In order to visually compare the difference of spectrograms of the original wideband speech, transmitted narrowband speech, and expanded wideband speech, adult male in Table 1 is chosen as an example and its spectrograms are shown in Figure 7a,b,c. From Figure 7c, we note that after the speech bandwidth extension by the proposed method, the 4 to 8 KHz frequency components have significantly increased by comparing with the transmitted narrowband speech in Figure 7b. It can be noticeable that since the synthetic wideband speech is filtered by a

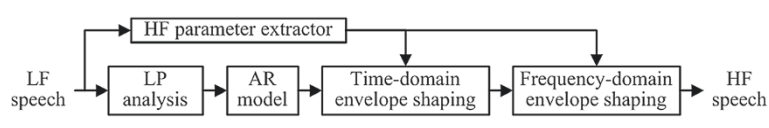

Figure 5 Block diagram of high-frequency speech restoration.

Table 1 Objective test results

Speech type	Distortion measure (dB)
Adult male	5.64
Adult female	5.82
Boy	5.51
Girl	5.42
Song	4.94

low-pass filter with 3.4 KHz stopband cutoff frequency (equivalent to 6.8 KHz after twice up-sampling), compared to Figure 7a, its spectrogram is evident in the dark at 7 to 8 KHz in Figure 7c.

It is self-evident that the watermark embedded into narrowband bit stream will decrease narrowband speech quality. Here, we use signal-to-noise ratio (SNR) of speech to evaluate the modified watermark method, whose results are shown in Table 2. We can find from Table 2 that SNR results of narrowband speech by the proposed watermark method are higher than the conventional LSB method.

3.2 Subjective evaluation

Subjective evaluation is to determine the speech quality by a person's hearing experience. Comparison mean opinion score (CMOS) method is used in this paper, and its scoring criteria is shown in Table 3.

There are four groups of wideband speech samples as subjective test set. The groups are labeled by female, male, boy, and girl, and each group has two different talkers. The length of each wideband (WB) speech sample is 8 s. Every person spoke five sentences, where one sentence is for pre-listening and other four sentences are for testing. The above four groups of test samples are coded-decoded with eight kinds of bit rates by adaptive multi-rate (AMR) codec and nine kinds of bit rates by AMR-WB codec respectively. The higher the coding rate is, the better the speech quality is. The same test samples are also coded-decoded by the proposed BWE method. The speech sample process is shown in Figure 8.

Because human auditory and subjective perceptions are based on personal experiences, knowledge background, test environment, and mental state, each person's subjective experience on the same speech will drift, but the difference is small. In order to make sure that the test situation can truly reflect the speech quality in the test, the 32 listeners (16 females and 16 males), whose ages are between 20 and 40, are invited for test experiments in the same test environment. None of the listeners had any hearing handicap, and they are native speakers of Chinese. The listeners have experience about communications facilities; especially, they were not engaged in

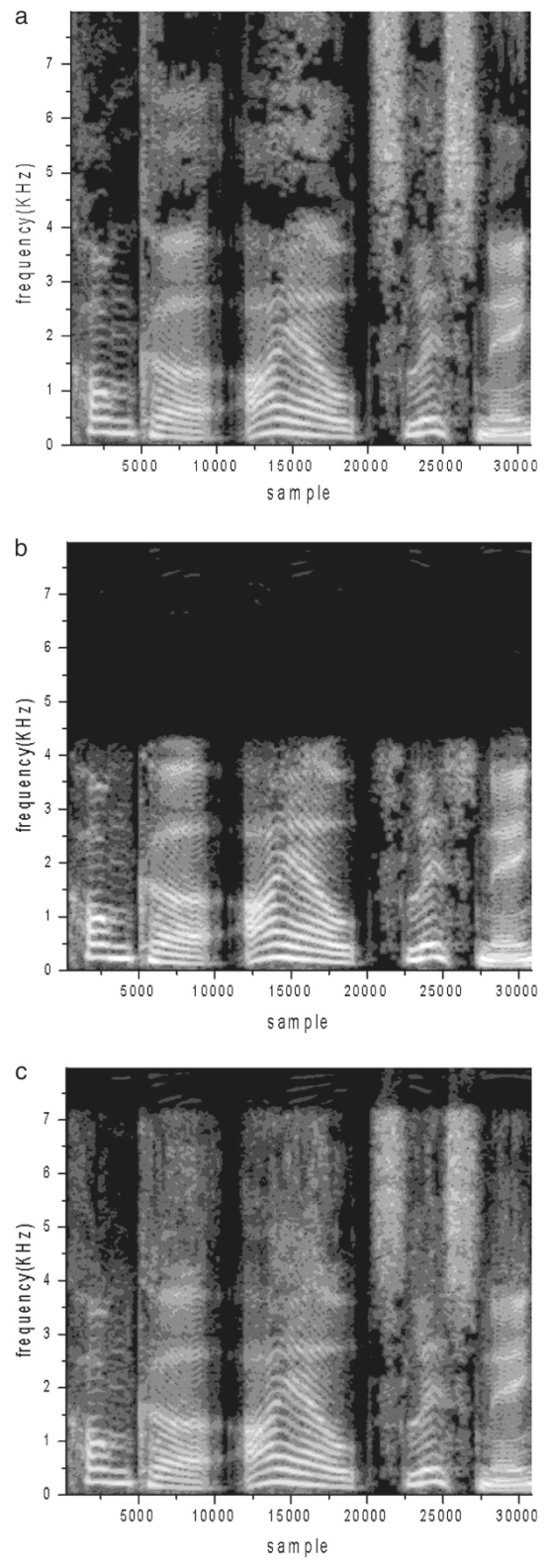

Figure 7 Comparison of the bandwidth extension spectrum. (**a**) Spectrum of original wideband speech. (**b**) Spectrum of transmitted narrowband speech. (**c**) Spectrum of wideband speech with proposed BWE method.

Table 2 Signal-to-noise ratio of narrowband speech

Method	SNR of narrowband speech after G.711 decode (dB)
Without watermark	36.92
LSB	31.58
Proposed	34.71

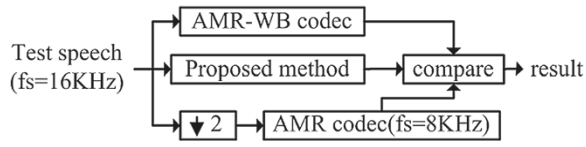

Figure 8 Block diagram of speech sample process.

communications or signal processing work and did not participate in any speech aspects of the subjective test in the recent 6 months.

Before formal listening tests, listeners was told of the main idea of the experiment. When the listeners understood the guidance, they will first listen to the initial situation and give their advices. Any technical problems, such as test principle or distortion degree, was forbidden before all experiments are over. In order to reduce the tiredness of the listeners, the test was divided into blocks. When test was ongoing, the listeners were not allowed to know the test results of other persons.

Figure 9 shows the distributions of subjective test among AMR 12.2 kbps, adaptive multi-rate-wideband (AMR-WB) 18.25 kbps and the proposed BWE method. In Figure 9, the average CMOS and its 95% confidence interval are also shown on the horizontal axis. Figure 9a shows the scores given in the comparison between the normal AMR codec at 12.2 kbps and the proposed BWE method. Figure 9b shows the scores given in the comparison between the AMR-WB codec at 18.25 kbps and the proposed BWE method. The black lines in abscissa in the Figure 9 represent the average scores in the test results. It can be seen from the Figure 9 that the average CMOS of the proposed method is slightly better than the AMR-WB codec at 18.25 kbps. However, compared with the results of AMR codec at 12.2 kbps, the performance of proposed method has greater improvement.

Most speech bandwidth extension methods are based on Gaussian mixture model or neural network model. In order to verify the effectiveness of proposed method, we made an experiment to compare the proposed method with references [5,6] by the CMOS. In the test, the 32 listeners (16 females and 16 males), whose ages are between 20 and 40, are invited for test experiments in the same test environment. None of the listeners had any hearing handicap, and they are native speakers of Chinese. After the experiment, the comparison result is shown in Table 4. We can see from the Table 4 that the average CMOS of the proposed method is slightly higher than that of reference [5], but compared with the reference [6], the proposed method has better performance.

Table 3 Signal to noise ratio of narrowband speech

Comparison	Score
A is much better than B	+3
A is better than B	+2
A is slightly better than B	+1
A is the same with B	0
A is slightly worse than B	−1
A is worse than B	−2
A is much worse than B	−3

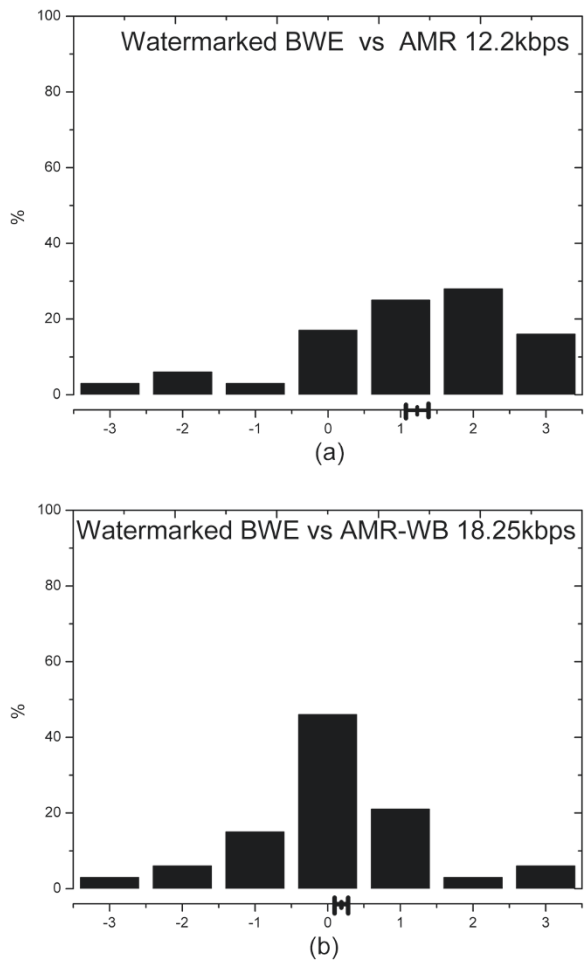

Figure 9 Distributions of the subjective test for different bit rates. (**a**) Watermarked BWE vs. AMR 12.2 kbps. (**b**) Watermarked BWE vs. AMR-WB 18.25 kbps.

Table 4 Comparison results of proposed method and ones by Pulakka et al. [5,6]

Method	CMOS	Confidence interval (%)
[5]	1.17	95
[6]	1.05	95
Proposed method	1.21	95

4 Conclusions

A speech bandwidth extension method based on the modified audio watermark is proposed in this paper. The high-frequency speech information as watermark is embedded in the narrowband (i.e., low-frequency) speech bit stream. A modified LSB watermark method based on the characteristics of the communication protocol and the human hearing perception is proposed and used in the proposed BWE method. The objective and subjective evaluations show that the quality of speech synthesized by the proposed method is better than narrowband speech and is comparable to AMR-WB codec at 18.25 kbps.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by National Natural Science Foundation of China (nos. 61172107, 61172110, and 60772161), Dalian Municipal Science and Technology Fund Scheme (no. 2008J23JH025), Specialized Research Fund for the Doctoral Program of Higher Education of China (no. 200801410015), and the Fundamental Research Funds for the Central Universities of China (no. DUT13LAB06).

Received: 12 February 2013 Accepted: 13 May 2013
Published: 6 June 2013

References

1. G. 711 ITU-T Recommendation, Pulse code modulation (PCM) of voice frequencies. (ITU-T, 1972)
2. MD Plumpe, TF Quatieri, DA Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. **7**(5), 569–586 (1999)
3. F Nagel, S Disch, S Wilde, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. A continuous modulated single sideband bandwidth extension, pp. 357–360. Texas, 14–19 March 2010
4. G Fuchs, R Lefebvre, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A new post-filtering for artificially replicated high-band in speech coders, pp. 713–716. Toulouse, 14–19 May, 2006
5. H Pulakka, U Remes, K Palomaki, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech bandwidth extension using gaussian mixture model-based estimation of the highband Mel spectrum, pp. 5100–5103. Prague, 22–27 May 2011
6. H Pulakka, P Alku, Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Mel spectrum. IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2170–2183 (2011)
7. TV Pham, F Schaefer, G Kubin, in *3th IEEE International Conference on Communications and Electronics (ICCE) Nha Trang*. A novel implementation of the spectral shaping approach for artificial bandwidth extension, pp. 262–267. 11–13 August 2010
8. P Bauer, T Fingscheidt, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An HMM-based artificial bandwidth extension evaluated by cross-language training and test, pp. 4589–4592. Las Vegas, 31 March–4 April 2008
9. Naofumi, A band extension technique for G.711 speech using steganography. IEICE Trans. Commun. **E89-B**(6), 1896–1898 (2006)
10. M Mohan, DB Karpur, M Narayan, in *IEEE International Conference on Communications and Signal Processing (ICCSP)*. Artificial bandwidth extension of narrowband speech using Gaussian mixture model, pp. 410–412. Kerala, 10–12 February 2011
11. S Chen, H Leung, in *IEEE International Symposium on Circuits and Systems (ISCAS)*. Artificial bandwidth extension of telephony speech by data hiding, pp. 3151–3154. Kobe, 23–26 May 2005
12. B Geiser, P Vary, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension, pp. 533–536. Honolulu, Hawaii, 15–20 April 2007
13. D Esteban, C Galand, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Application of quadrature mirror filters to split band voice coding schemes, pp. 191–195. Hartford, May 1977
14. ITU-T Recommendation G.729.1: G.729-based embedded variable bit-rate coder: an 8–32 kbit/s scalable wideband coder bit stream interoperable with G.729, (ITU-T, 2006)
15. T Nomura, M Iwadare, M Serizawa, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A bitrate and bandwidth scalable CELP coder, pp. 341–344. Seattle, 12–15 May 1998
16. F Mustiere, M Bouchard, M Bolic, in *Canadian Conference on Electrical and Computer Engineering*. Bandwidth extension for speech enhancement, pp. 1–4. Calgary, 2–5 May 2010
17. P Jax, P Vary, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. An upper bound on the quality of artificial bandwidth extension of narrowband speech signals, pp. 237–240. Orlando, 13–17 May 2002
18. HW Hsu, CM Liu, Decimation-whitening filter in spectral band replication. IEEE Trans. Audio, Speech, Lang Process. **19**(8), 2304–2313 (2011)
19. J Zhang, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Bandwidth extension for China AVS-M standard, pp. 4149–4152. Taipei, 19–24 April 2009

doi:10.1186/1687-4722-2013-10
Cite this article as: Chen *et al.*: **An audio watermark-based speech bandwidth extension method.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:10.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com