Circuits, Systems, and Signal Processing (2022) 41:2290–2307
https://doi.org/10.1007/s00034-021-01893-5

# Speech Bandwidth Extension Using Data Hiding Based on Discrete Hartley Transform Domain

**Yuya Hosoda**[1] · **Arata Kawamura**[2] · **Youji Iiguni**[1]

Received: 18 June 2020 / Revised: 19 October 2021 / Accepted: 20 October 2021 /
Published online: 18 November 2021


## Abstract

The public switching telephone network restricts speech signals to a narrow bandwidth (NB) of 0.3–3.4 kHz, which results in a perceived reduction in quality due to the missing upper bandwidth (UB) spectrum of 3.4–7 kHz. This paper proposes a speech bandwidth extension method that reconstructs the missing UB spectrum using side information. A sender side obtains a UB spectral envelope and relative gains between NB and UB excitation signals as side information. Side information is then converted into a binary signal using two codebooks. Using speech steganography based on the discrete Hartley transform (DHT) domain, the proposed method robustly embeds the binary signal into an amplitude spectrum of the NB speech signal in the high-frequency bandwidth of 3.4–4.6 kHz to produce a composite narrow bandwidth (CNB) speech signal. On a receiver side, the missing UB spectrum is reconstructed using side information extracted from the CNB speech signal. Theoretical and simulation analysis shows that side information is retrieved from the CNB speech signal accurately. Subjective listening tests and objective measures also show that the proposed method enhances the quality of the NB speech signal by reconstructing the missing UB spectrum.

**Keywords** Speech bandwidth extension · Speech steganography · Discrete Hartley transform · Speech quality

✉ Yuya Hosoda
hosoda@sip.sys.es.osaka-u.ac.jp

Arata Kawamura
kawamura@cc.kyoto-su.ac.jp

Youji Iiguni
iiguni@sys.es.osaka-u.ac.jp

[1] Graduate School of Engineering Science, Osaka University, Osaka, Japan

[2] Faculty of Information Science and Engineering, Kyoto Sangyo University, Kyoto, Japan

## 1 Introduction

The public switching telephone network uses speech codecs for low latency communication. It is desirable to utilize wide bandwidth (WB) speech codecs, such as AMR-WB [17], which has been standardized to improve quality significantly. However, the old analog telephone system supports only the narrow bandwidth (NB) speech codecs, such as G.711 [14] and G.729 [18]. NB speech signals are limited in the bandwidth of 0.3–3.4 kHz, resulting in a perceived reduction in quality compared to WB speech signals with the bandwidth of 0.3–7 kHz [40]. The old analog telephone system renovation also takes much effort for both the sender and receiver sides [22]. To enhance the quality of the NB speech signal, speech enhancement approaches have received much attention.

A speech bandwidth extension (BWE) method is a speech enhancement approach that reconstructs the missing upper bandwidth (UB) spectrum of 3.4–7 kHz using a source-filter model of speech production [21]. The source-filter model represents a speech signal by a convolution of an excitation signal and a spectral envelope [24]. A UB excitation signal is generated from the existing NB excitation signal by frequency shift [25] or noise modulation [35]. A UB spectral envelope is estimated using codebooks [38] based on statistical models, such as Gaussian mixture models [34] and neural networks [1]. However, BWE methods have the performance limitation to reconstruct the missing UB spectrum. Jax et al. showed that the maximum achievable performance of the UB spectral envelope estimation depends on mutual information (MI) between NB and UB spectra and that it is relatively low to reconstruct the missing UB spectrum [20]. This is because an NB spectrum has a one-to-many relationship with UB spectra [2]. Nilsson et al. also demonstrated that MI of the consonants, especially fricatives, is lower than that of vowels [26].

A solution to resolve the performance limitation is to transmit side information about the missing UB spectrum. However, transmitting both side information and an NB speech signal may cause high latency communication due to the increased amount of information. Researchers have devised BWE methods that transmit side information using speech steganography without increasing the amount of information [4–7,12,28–30,37,39]. Speech steganography embeds side information into a hidden channel of the NB speech signal to generate a composite narrow bandwidth (CNB) speech signal. When the receiver side supports speech steganography, the missing UB spectrum can be reconstructed using side information extracted from the CNB speech. Otherwise, the received CNB speech signal is used directly as an NB speech signal. Therefore, BWE methods using speech steganography need to minimize the quality decline of the CNB speech signals due to embedding side information.

A BWE method using bitstream data hiding embeds side information into a bitstream of the encoded NB speech signal [7]. BWE methods using a joint coding technique also incorporate embedding side information into encoding NB speech signals [28,39]. Although the quality of the CNB speech signal is equivalent to that of the NB speech signal, these methods treat with only a specific NB speech codec. To correspond to various NB speech codecs, BWE methods using signal domain speech steganography have been devised, which embeds side information into an NB speech

signal before encoding [4–6,12,29,30,37]. As side information, various feature vectors have been utilized to reconstruct the missing UB spectrum.

Prasad et al. adopted code excited linear prediction (CELP) parameters as feature vectors [30]. Since CELP parameters generally have covered NB speech signals, this method should set specified CELP parameters for UB speech signals. Methods using the part of the UB power spectra have also been devised [6,12]. The most straightforward approach is to utilize a relative gain between NB and UB excitation signals and line spectral frequencies (LSF), representing a UB spectral envelope [4,5,29,37]. A relative gain is required to avoid an overestimation of the power of the UB speech signal [3,27]. A sender side converts feature vectors into a binary signal using a codebook. The binary signal is then embedded into an NB speech signal using signal domain speech steganography based on some transform domains.

Chen et al. embed a binary signal into an NB speech signal using dither quantization in the time domain [4,5]. Although dither quantization requires less processing time, a bit error for the binary signal occurs due to artifacts such as speech codecs and noise through the telephone system. Sagi et al. embedded a binary signal using the scalar Costa scheme in the discrete Hartley transform (DHT) domain [37]. While it is robust against artifacts, the capacity of binary signals to be embedded depends on the power of the NB speech signal.

Prasad et al. proposed a BWE method using transform-domain data hiding (TDDH) based on the discrete Fourier transform (DFT) domain [29]. TDDH is robust signal domain speech steganography that converts a binary signal into a hidden vector and embeds it into a magnitude spectrum in the high-frequency bandwidth of 3.4-4 kHz, not depending on the power of the NB speech signal. Since human hearing is little sensitive to the distortion of the magnitude spectrum in the high-frequency bandwidth due to TDDH [31,32], the quality of the CNB speech signal is almost equivalent to that of the NB speech signal. However, the hidden vector may contain negative values despite that the magnitude spectrum accepts only nonnegative values. An offset is thus required to be set to embed a hidden vector into a magnitude spectrum, which degrades the quality of the CNB speech signal. Besides, a UB speech signal is generated with non-overlapping, which results in the discontinuity between frames. Also, it is not easy to reproduce the slight UB sound pressure change by calculating a relative gain in each frame.

In this paper, we propose a BWE method using TDDH based on the DHT domain. The proposed method has two advantages. First, we avoid setting the offset by embedding a hidden vector into an amplitude spectrum of the NB speech signal in the DHT domain where negative values can be accepted. Furthermore, the conventional method [29] embeds a common hidden vector into the bandwidths of 3.4–4 kHz and 4–4.6 kHz in the DFT domain with symmetry at a Nyquist frequency, while the proposed method embeds different hidden vectors because the DHT domain is asymmetric, which improves the robustness against artifacts by embedding the hidden vector into the wider bandwidth. Second, the proposed method generates a UB speech signal with overlapping to avoid the discontinuity between frames. Here, relative gains are calculated in sub-frames to reproduce the slight UB sound pressure change over time.

This paper is organized as follows: In Sect. 2, we present a method of generating a CNB speech signal. Section 3 describes a BWE method using side information

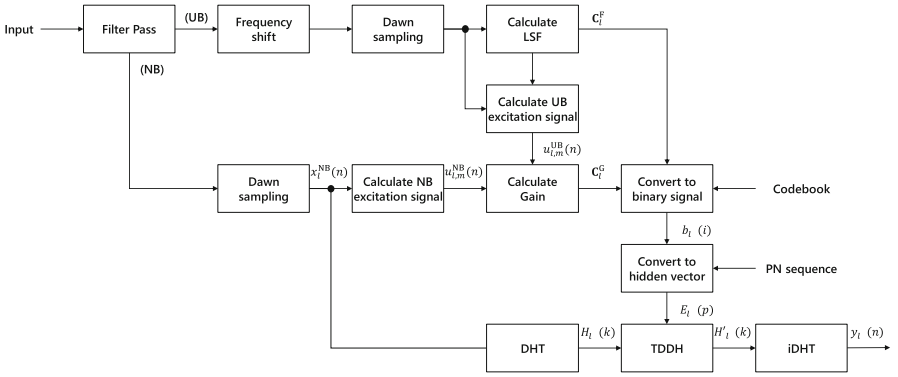


**Fig. 1** Block diagram of the CNB speech signal generation

extracted from the CNB speech signal. We analyze the performance of TDDH based on the DHT domain in Sect. 4. Subjective listening tests and objective measures for the proposed method are discussed in Sect. 5. Finally, conclusions are given in Sect. 6.

## 2 Composite Narrow Bandwidth Speech Signal Generation

Figure 1 shows a block diagram of the CNB speech signal generation. First, an input speech signal is separated into NB and UB speech signals using a band-pass filter with cutoff frequencies of 0.3 kHz and 3.4 kHz and a high-pass filter with a cutoff frequency of 3.4 kHz. To reduce a redundancy, the UB speech signal is frequency-shifted and down-sampled. Feature vectors of LSF and relative gains are extracted from the frequency-shifted and down-sampled UB speech signal and then quantized into a binary signal using two codebooks. The binary signal is converted into a hidden vector using the spread spectrum scheme with a pseudo-noise (PN) sequence to enhance the robustness against artifacts. Finally, the hidden vector is embedded into an amplitude spectrum in the high-frequency bandwidth of 3.4–4.6 kHz using TDDH based on the DHT domain, and a CNB speech signal is generated by inverse DHT (iDHT).

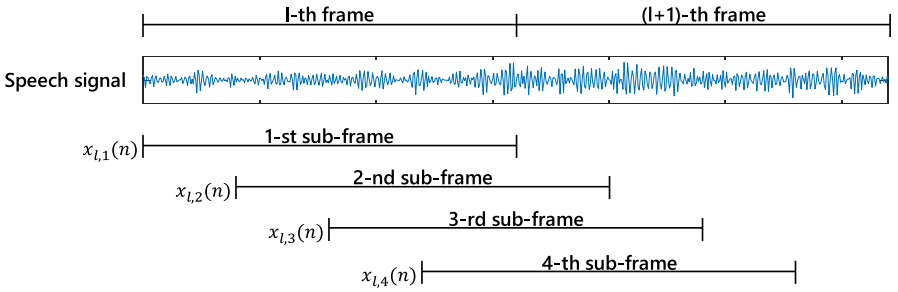


**Fig. 2** Sub-frame definition for the relative gain calculation

The proposed method calculates LSF of the UB speech signal with non-overlapping. First, autoregressive (AR) coefficients in $l$-th frame $a_l(j)$ $(j = 1, \ldots, J)$ are given by solving the following equation using the Levinson–Durbin algorithm:

$$\sum_{j=1}^{J} a_l(j) r_l(|q - j|) = r_l(q), \ q = 1, \ldots, J, \tag{1}$$

where $r_l(q)$ and $J$ denote a modified autocorrelation coefficient and the order of the AR coefficients, respectively. The AR coefficients are then converted into LSF $F_l(j)$ to suppress the quantization error [13].

Next, the proposed method calculates a relative gain between NB and UB excitation signals in each sub-frame. Figure 2 shows the sub-frame definition for the relative gain calculation, where we consider that a UB speech signal is generated with 75% overlapping at the receiver side in this paper. Let $x_{l,m}(n)$ $(n = 0, \ldots, N-1)$ denote a speech signal at $m$-th sub-frame, where $N$ denotes the number of frame samples. An excitation signal $u_{l,m}(n)$ is defined as:

$$u_{l,m}(n) = x_{l,m}(n) - \sum_{j=1}^{J} a_l(j) x_{l,m}(|n - j|). \tag{2}$$

Let $u_{l,m}^{\mathrm{NB}}(n)$ and $u_{l,m}^{\mathrm{UB}}(n)$ denote NB and UB excitation signals, respectively. The proposed method calculates a relative gain $G_{l,m}$, following as

$$G_{l,m} = 20 \log_{10} \left\{ \sum_{n=0}^{N-1} \left( u_{l,m}^{\mathrm{UB}}(n) \right)^2 \right\} - 20 \log_{10} \left\{ \sum_{n=0}^{N-1} \left( u_{l,m}^{\mathrm{NB}}(n) \right)^2 \right\}. \tag{3}$$

Finally, we obtain feature vectors for LSF $\mathbf{C}_l^{\mathrm{F}} = [F_l(1), \ldots, F_l(J)]^{\mathrm{T}}$ and relative gains $\mathbf{C}_l^{\mathrm{G}} = [G_{l,1}, \ G_{l,2}, \ G_{l,3}, \ G_{l,4}]^{\mathrm{T}}$, where $[\cdot]^{\mathrm{T}}$ denotes a transpose operation.

When feature vectors are grouped and converted into a binary signal using a codebook, the length of the binary signal needs to be increased to suppress the quantization error [23]. Nevertheless, as the length of the binary signal increases, the quality decline of the CNB speech signal due to TDDH becomes more serious [32]. Hence, the proposed method converts each feature vector into binary signals separately using some codebooks, as well as G.729 [18]. Let $2^{N^{\mathrm{F}}}$ and $2^{N^{\mathrm{G}}}$ $(N^{\mathrm{F}}, \ N^{\mathrm{G}} > 0)$ denote the size of the codebook for LSF and relative gains, respectively. The feature vectors are quantized with $N^{\mathrm{F}}$ and $N^{\mathrm{G}}$ binary digits, respectively. In this paper, we obtain codebooks by the Linde–Buzo–Gray training algorithm [23]. The proposed method generates a binary signal $b_l(i)(\in \{1, \ -1\}, i = 0, 1, \ldots, N^{\mathrm{S}} - 1)$ by combining these binary signals, where $N^{\mathrm{S}}$ denotes the total bit length such as $N^{\mathrm{S}} = N^{\mathrm{F}} + N^{\mathrm{G}}$. In this case, a synchronization sequence such as $111 \ldots 11$ has been prepared to accomplish the frame synchronization between the sender and receiver sides [10].

To enhance the robustness against artifacts, the proposed method converts the binary signal to a hidden vector using the spread spectrum scheme [8]. Let $P$ denote the length of the bandwidth where the hidden vector is embedded into an amplitude spectrum.

With a PN sequence $Q(p,\ i)(\in \{1,\ -1\},\ p = 0, 1, \ldots, P-1)$, we obtain a hidden vector

$$E_l(p) = \beta \cdot \sum_{i=0}^{N_S-1} Q(p,\ i) b_l(i), \tag{4}$$

where $\beta(>\ 0)$ denotes the strength of TDDH. The larger $\beta$, the more robust the hidden vector against artifacts, but the more significantly the quality decline of the CNB speech signal due to TDDH, and vice versa. Also, as shown in Sect. 4, we set $\beta$ as a positive value to avoid reversing the binary signal extracted from the CNB speech signal. In this paper, we empirically fix at $\beta = 0.01$. Besides, we generate a PN sequence using Hadamard codes [9].

The proposed method embeds the hidden vector into an amplitude spectrum in the high-frequency bandwidth of 3.4–4.6 kHz with non-overlapping. Let $x_l^{\mathrm{NB}}(n)$ denote an NB speech signal. We define an amplitude spectrum $H_l(k)(k = 0, 1, \ldots, N-1)$ as

$$H_l(k) = \sum_{n=0}^{N-1} x_l^{\mathrm{NB}}(n) \mathrm{cas}\left(\frac{2\pi nk}{N}\right), \tag{5}$$

with

$$\mathrm{cas}(t) = \cos(t) + \sin(t). \tag{6}$$

The proposed method then embeds the hidden vector into the amplitude spectrum, following as

$$H_l'(k) = \begin{cases} H_l(k), & k = 0, \ldots, (N-P)/2 - 1 \\ E_l\big(k - (N-P)/2\big), & k = (N-P)/2, \ldots, (N+P)/2 - 1 \\ H_l(k), & k = (N+P)/2, \ldots, N-1 \end{cases}. \tag{7}$$

Finally, we obtain a CNB speech signal $y_l(n)$, following as

$$y_l(n) = \frac{1}{N} \sum_{k=0}^{N-1} H'_l(k) \mathrm{cas}\left(\frac{2\pi nk}{N}\right). \tag{8}$$

## 3 Speech Bandwidth Extension Using Side Information Extracted from CNB Speech Signal

Figure 3 shows a block diagram of the BWE method using side information extracted from a CNB speech signal. First, a hidden vector is extracted from the amplitude spectrum of the received CNB speech. Here, an NB speech signal is generated by iDHT from the amplitude spectrum, where the hidden vector has been extracted. The

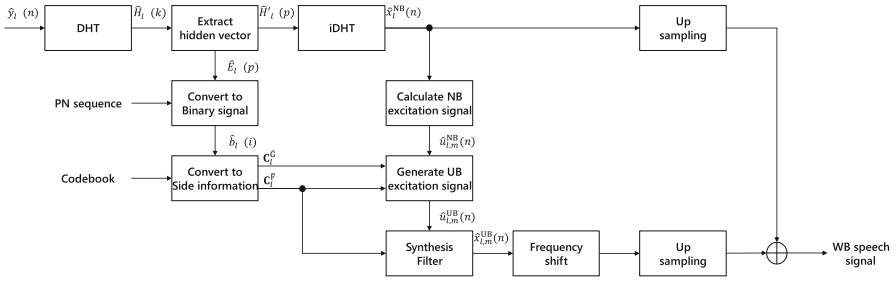


**Fig. 3** Block diagram of the BWE method using side information extracted from CNB speech signal

hidden vector is converted into a binary signal using a PN sequence. Feature vectors for LSF and relative gains are retrieved from the binary signal using codebooks. A UB excitation signal is generated from the NB excitation signal and the retrieved relative gains, and a UB spectral envelope is obtained from the retrieved LSF. The proposed method calculates the UB speech signal with overlapping to avoid the discontinuity between frames. Finally, the frame-shifted and up-sampled UB speech signal is added to the up-sampled NB speech signal to generate a WB speech signal.

Let $\hat{H}_l(k)$ be an amplitude spectrum of the received CNB speech. An hidden vector $\hat{E}_l(p)$ is extracted, following as

$$\hat{E}_l(p) = \hat{H}_l\big((N-P)/2 + p\big). \tag{9}$$

The extracted hidden vector is then converted into a binary signal

$$\hat{b}_l(i) = \mathrm{sgn}\left[\sum_{p=0}^{P-1} Q(p,\ i)\hat{E}_l(p)\right], \tag{10}$$

where sgn[·] denotes a sign function. The retrieved feature vectors for LSF ($\mathbf{C}_l^{\hat{F}} = [\hat{F}_l(1), \ldots, \hat{F}_l(J)]^{\mathrm{T}}$) and relative gains $\mathbf{C}_l^{\hat{G}} = [\hat{G}_{l,1},\ \hat{G}_{l,2},\ \hat{G}_{l,3},\ \hat{G}_{l,4}]^{\mathrm{T}}$ are obtained from the binary signal using codebooks. The proposed method also reuses the CNB speech signal as an NB speech signal. Let $\hat{H'}_l(k)$ denote an amplitude spectrum where the hidden vector has been extracted:

$$\hat{H'}_l(k) = \begin{cases} \hat{H}_l(k), & k = 0, \ldots, (N-P)/2 - 1 \\ 0, & k = (N-P)/2, \ldots, (N+P)/2 - 1 \\ \hat{H}_l(k), & k = (N+P)/2, \ldots, N-1 \end{cases}. \tag{11}$$

An NB speech signal is calculated from the amplitude spectrum using iDHT, following as

$$\hat{x}_l^{\mathrm{NB}}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{H'}_l(k)\mathrm{cas}\left(\frac{2\pi nk}{N}\right). \tag{12}$$

With $\mathbf{C}_l^{\hat{G}}$, the proposed method generates a UB excitation signal $\hat{u}_{l,m}^{\mathrm{UB}}(n)$, following as

$$\hat{u}_{l,m}^{\mathrm{UB}}(n) = \sqrt{10^{\frac{\hat{G}_{l,m}}{20}}}\,\hat{u}_{l,m}^{\mathrm{NB}}(n), \tag{13}$$

where $\hat{u}_{l,m}^{\mathrm{NB}}(n)$ denotes an NB excitation signal. Let $\hat{a}_l(j)$ denote AR coefficients converted from $\hat{F}_l(j)$. We generate a UB speech signal

$$\hat{x}_{l,m}^{\mathrm{UB}}(n) = \hat{u}_{l,m}^{\mathrm{UB}}(n) + \sum_{j=1}^{J} \hat{a}_l(j)\hat{x}_{l,m}^{\mathrm{UB}}(|n-j|). \tag{14}$$

Finally, we obtain a WB speech signal by adding the up-sampled NB speech signal into the frame-shifted and up-sampled UB speech signal with overlapping.

## 4 Performance Analysis for Transform-Domain Data Hiding Based on Discrete Hartley Transform Domain

This section discusses the performance analysis of TDDH based on the DHT domain. We assume that a CNB speech signal suffers artifacts based on the additive white Gaussian noise (AWGN). The received CNB speech signal $\hat{y}_l(n)$ is written as:

$$\hat{y}_l(n) = y_l(n) + g(n), \tag{15}$$

where $g(n)$ denotes white Gaussian noise. On the DHT domain, Eq. (15) is interpreted as:

$$\hat{H}_l(k) = H_l'(k) + J(k), \tag{16}$$

where $J(k)$ denotes an amplitude spectrum of $g(n)$. By substituting Eqs. (7) and (16) into Eq. (9), a relation equation is given as:

$$\hat{E}_l(p) = E_l(p) + J'(p), \tag{17}$$

with

$$J'(p) = J\big((N'-P)/2 + p\big). \tag{18}$$

By substituting Eqs. (4) and (17) into Eq. (10), Eq. (10) is also interpreted as

$$\hat{b}_l(i) = \operatorname{sgn}\left[\sum_{p=0}^{P-1} Q(p,\ i)\left(E_l(p) + J'(p)\right)\right]$$
$$= \operatorname{sgn}\left[\sum_{p=0}^{P-1} (\beta Q(p,\ i)Q(p,\ i)b_l(i) + \sum_{i'\neq i} \beta Q(p,\ i)Q(p,\ i')b_l(i') + Q(p,\ i)J'(p)\right)\right]. \tag{19}$$

Note that a PN sequence is orthogonal such that $\sum_{p=0}^{P-1} Q(p,\ i)Q(p,\ i') = P \cdot \delta_{i-i'}$, where $\delta_i$ is the Kronecker delta. Equation (19) is thus rewritten as

$$\hat{b}_l(i) = \operatorname{sgn}\left[\beta P b_l(i) + \sum_{p=0}^{P-1} Q(p,\ i)J'(p)\right]. \tag{20}$$

In a clean environment with $J'(p) = 0$, we have $\hat{b}_l(i) = \operatorname{sgn}[\beta P b_l(i)]$. Because of $P > 0$, the equation is then rewritten as $\hat{b}_l(i) = \operatorname{sgn}[\beta b_l(i)]$. If $\beta < 0$, we have $\hat{b}_l(i) \neq b_l(i)$. Hence, we set $\beta > 0$. A bit error occurs in the case such that $\operatorname{sgn}[b_l(i)] \cdot \operatorname{sgn}\left[\sum_{p=0}^{P-1} Q(p,\ i)J'(p)\right] = -1$ and $|\beta P b_l(i)| \leq |\sum_{p=0}^{P-1} Q(p,\ i)J'(p)|$.

We define a variable $\hat{d}_l(i)$ from Eq. (20) as

$$\hat{d}_l(i) = \beta P b_l(i) + \sum_{p=0}^{P-1} Q(p,\ i)J'(p). \tag{21}$$

According to the central limit theorem [11], the conditional probability distribution $f(\hat{d}_l(i) \mid b_l(i))$ is given as:

$$f(\hat{d}_l(i) \mid b_l(i) = 1) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{Q}}^2}} e^{-\frac{(\hat{d}_l(i) - \beta P)^2}{2\sigma_{\mathrm{Q}}^2}}, \tag{22}$$

$$f(\hat{d}_l(i) \mid b_l(i) = -1) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{Q}}^2}} e^{-\frac{(\hat{d}_l(i) + \beta P)^2}{2\sigma_{\mathrm{Q}}^2}}, \tag{23}$$

where $\sigma_{\mathrm{Q}}^2$ denotes the variance of the variable $\sum_{p=0}^{P-1} Q(p,\ i)J'(p)$. We then transform $\sigma_{\mathrm{Q}}^2$ as

$$\begin{aligned}\sigma_{\mathrm{Q}}^2 &= \mathrm{E}\left[\left(\sum_{p=0}^{P-1} Q(p,\ i)J'(p)\right)^2\right] \\ &= \mathrm{E}\left[\sum_{p=0}^{P-1}\sum_{p'=0}^{P-1} Q(p,\ i)Q(p',\ i)J'(p)J'(p')\right] \\ &= \mathrm{E}\left[\sum_{p=0}^{P-1}\sum_{p'=0}^{P-1} N^{\mathrm{S}}\cdot\delta_{p-p'}J'(p)J'(p')\right] \\ &= N^{\mathrm{S}}\cdot\sum_{p=0}^{P-1}\mathrm{E}\left[J'(p)^2\right] \\ &= N^{\mathrm{S}}P\sigma_{\mathrm{J}}^2 \end{aligned} \tag{24}$$

where $\sigma_{\mathrm{J}}^2$ denotes the variance of $J'(p)$. In the case of $\hat{d}_l > 0$, $\hat{b}_l(i) = 1$ for Eq. (20). The conditional probability for $p(\hat{b}_l(i) = 1 \mid b_l(i) = -1)$ is thus given as:

$$\begin{aligned} p(\hat{b}_l(i) = 1 \mid b_l(i) = -1) &= \int_0^\infty f(\hat{d}_l(i) \mid b_l(i) = -1)\mathrm{d}\hat{d} \\ &= \frac{1}{\sqrt{2\pi\sigma_{\mathrm{Q}}^2}}\int_0^\infty e^{-\frac{(\hat{d}_l(i)+\beta P)^2}{2\sigma_{\mathrm{Q}}^2}}\mathrm{d}\hat{d} \\ &= \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta^2P^2}{2\sigma_{\mathrm{Q}}^2}}\right) \\ &= \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta^2P}{2N^{\mathrm{S}}\sigma_{\mathrm{J}}^2}}\right), \end{aligned} \tag{25}$$

where $\mathrm{erfc}(q) = \frac{2}{\sqrt{\pi}}\int_q^\infty e^{-t^2}\mathrm{d}t$ denotes a complementary error function. Similarly, the conditional probability $p(\hat{b}_l(i) = -1 \mid b_l(i) = 1)$ is given as:

$$p(\hat{b}_l(i) = -1 \mid b_l(i) = 1) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta^2P}{2N^{\mathrm{S}}\sigma_{\mathrm{J}}^2}}\right). \tag{26}$$

We assume that the prior probabilities are equiprobable such as $p(b_l(i) = 1) = p(b_l(i) = -1) = 1/2$, which has been used for the bit error calculation [29,41]. Based on Eqs. (25) and (26), we calculate the probability for the bit error

$$\begin{aligned} e_l = {} & p\big(\hat{b}_l(i) = -1 \mid b_l(i) = 1\big)\cdot p\big(b_l(i) = 1\big) \\ & + p\big(\hat{b}_l(i) = 1 \mid b_l(i) = -1\big)\cdot p\big(b_l(i) = -1\big) \end{aligned}$$

$$= \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta^2 P}{2N^{\mathrm{S}}{\sigma_{\mathrm{J}}}^2}}\right). \tag{27}$$

We find that $e_l$ decreases as $P$ increases. That is, TDDH based on the DHT domain improves the robustness against artifacts by embedding the hidden vector into the amplitude spectrum in the wider high-frequency bandwidth.

## 5 Subjective Listening Tests and Objective Measures

This section describes subjective listening tests and objective measures for the proposed method. First, we verified the quality difference between NB and CNB speech signals. Second, we evaluated the quality of a generated WB speech signal where the missing UB spectrum has been reconstructed using side information extracted from the CNB speech signal. Besides, we verified the robustness against artifacts. In this paper, we assumed noise environments: AWGN at several signal-to-noise ratio (SNR) levels without or with speech codecs G.711 [14] and G.729 [18].

We used speech datasets taken from English speech corpus PTDB-TUG [33] and Japanese speech corpus ASJ-JNAS [19]. Speech samples taken from PTDB-TUG were used for training codebooks. Hundred speech samples taken from ASJ-JNAS were used for the performance analysis tests. A sampling rate for speech signals was 16 kHz. We adopted 10-order LSF ($J = 10$) to represent a UB spectral envelope using the Hamming window. Also, the number of frame samples was $N = 160$ (20 ms). Feature vectors were converted into a binary signal of $N^{\mathrm{S}} = 12$, where the proposed method utilized two codebooks of $N^{\mathrm{F}} = 8$ and $N^{\mathrm{G}} = 4$.

The proposed BWE method using TDDH based on the DHT domain with relative gains (BWE-HG) is compared with three different methods: a BWE method using TDDH based on the DFT domain with a relative gain (BWE-F) [29], a BWE method using TDDH based on the DHT domain with a relative gain (BWE-H), and a BWE method using TDDH based on the DFT domain with relative gains (BWE-FG). BWE-F and BWE-FG converted a binary signal into a hidden vector of $P = 12$ and embedded it into a magnitude spectrum in the high-frequency bandwidth of 3.4–4 kHz, where the common hidden vector was embedded in the bandwidth of 4–4.6 kHz because of the symmetry at a Nyquist frequency. Here, an offset was required to embed a hidden vector with negative values into a magnitude spectrum with non-negative values. For BWE-F and BWE-FG, Eqs. (4) and (10) are rewritten as:

$$E_l(p) = \beta \cdot \sum_{i=0}^{N_S-1} Q(p,\ i) b_l(i) + \beta P, \tag{28}$$

and

$$\hat{b}_l(i) = \mathrm{sgn}\left[\sum_{p=0}^{P-1} Q(p,\ i)\left(\hat{E}_l(p) - \beta P\right)\right], \tag{29}$$

**Table 1** Category for DMOS

| Score | Category |
|---|---|
| 1 | Degradation is very annoying |
| 2 | Degradation is annoying |
| 3 | Degradation is slightly annoying |
| 4 | Degradation is audible but not annoying |
| 5 | Degradation is inaudible |

**Table 2** Category for MOS

| Score | Category |
|---|---|
| 1 | Bad |
| 2 | Poor |
| 3 | Fair |
| 4 | Good |
| 5 | Excellent |

respectively. BWE-H and BWE-HG also converted a hidden vector of $P = 24$ and embedded it into an amplitude spectrum in high-frequency bandwidth of 3.4–4.6 kHz. While BWE-F and BWE-H calculated a relative gain in each frame and generated a UB speech signal with non-overlapping, BWE-FG and BWE-HG calculated relative gains in sub-frames and generated a UB speech signal with 75% overlapping. Here, BWE-F and BWE-H grouped feature vectors such as $\mathbf{C}'_l = [F_l(1), \ldots, F_l(J), G_{l,1}]^{\mathrm{T}}$ and converted it to a binary signal of $N^{\mathrm{S}} = 12$.

In subjective listening tests, 9 Japanese listeners between the age of 22 and 24 participated, who had normal hearing and have not trained before. A listener listened to test speech samples generated from two male and two female speakers through a headphone (MDR-7506) in a quiet room. A degradation category rating test [15] was employed to evaluate a CNB speech signal in comparison with an NB speech signal based on degradation mean opinion score (DMOS) in Table 1. An absolute category rating test [15] was also employed to evaluate the generated WB speech signals based on mean opinion score (MOS) in Table 2.

In objective quality measurements, we evaluated the perceptual transparency of the CNB speech signal using NB-PESQ [16]. NB-PESQ returned a score from -0.5 to 4.5. We also evaluated the perceptual similarity between original and estimated WB speech signals using log-spectral distance (LSD) [36]. In this paper, we define LSD as:

$$\mathrm{LSD} = \frac{1}{L}\sum_{l=0}^{L-1}\sqrt{\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\left[10\log_{10}\frac{P_l(k)}{\hat{P}_l(k)}\right]^2}, \tag{30}$$

where $P_l(k)$ and $\hat{P}_l(k)$ denote power spectra of the original and generated WB speech signals, respectively. Also, $\mathcal{K}$ is the set of the frequency indices at the bandwidth to

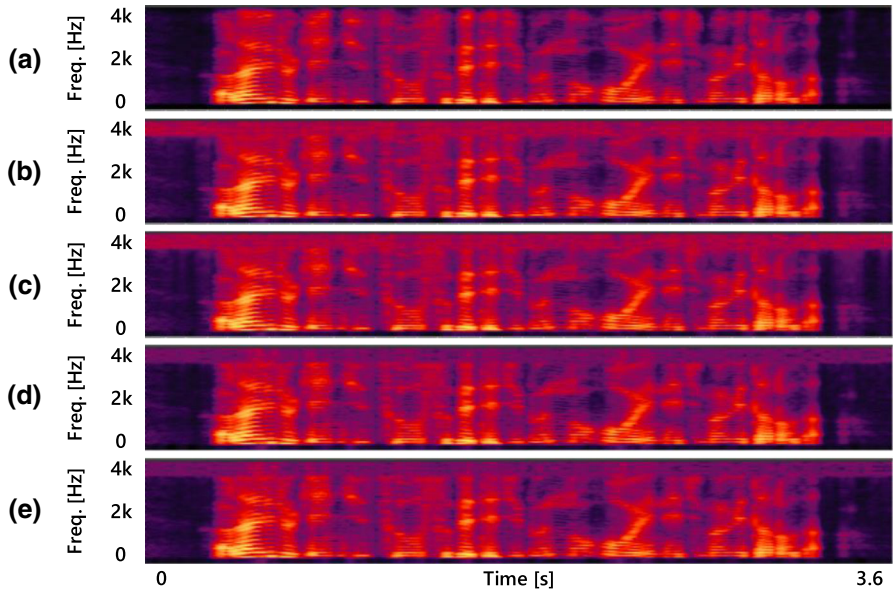


**Fig. 4** Sound spectrograms of NB and CNB speech signals. **a** NB speech signal. **b** CNB speech signal for BWE-F. **c** CNB speech signal for BWE-FG. **d** CNB speech signal for BWE-H. **e** CNB speech signal for BWE-HG

be analyzed, and $L$ denotes the number of analyzed frames. We utilized the Hamming window of 32 ms and analyzed the bandwidth of 3.4–7 kHz.

The first experiment verified the quality difference between NB and CNB speech signals. Figure 4 shows sound spectrograms for NB and CNB speech signals. Compared to the NB speech signal, the CNB speech signals had the spectral distortion in the high-frequency bandwidth of 3.4–4 kHz due to TDDH. It was also seen that BWE-F and BWE-FG had more serious spectral distortion because of the offset. Table 3 shows results of DMOS and NB-PESQ. Compared to DMOS for the CNB speech signal using TDDH based on the DFT domain (BWE-F and BWE-FG), NB-PESQ for the CNB speech signal using TDDH based on the DHT domain (BWE-H and BWE-HG) was higher by more than 0.30 points because of the needless of the offset. Also, DMOS for BWE-HG was over 3.60 because the proposed method avoided the discontinuity between frames by generating a UB speech signal with 75% overlapping. The proposed method therefore suppressed the quality decline due to TDDH in comparison with the conventional method [29].

The second experiment verified the quality of the generated WB speech signal. Figure 5 shows sound spectrograms of original and generated WB speech signals. It can be seen that the missing UB spectrum has been reconstructed successfully on the generated WB speech signals. Also, since BWE-F and BWE-H, and BWE-FG and BWE-HG have reconstructed the UB spectrum using a common feature vector, the sound spectrograms of these generated WB speech signals were identical, respectively. Compared to the method of generating a UB speech signal with a relative gain (BWE-F and BWE-H), the method of generating a UB speech signal with relative gains (BWE-

**Table 3** Subjective and objective quality assessments of CNB speech signals

| | | BWE-F | BWE-FG | BWE-H | BWE-HG |
|---|---|---|---|---|---|
| DMOS | Female A | 2.44 | 2.44 | 3.78 | 3.33 |
| | Female B | 2.67 | 2.11 | 3.56 | 3.56 |
| | Male A | 3.00 | 3.33 | 3.78 | 4.11 |
| | Male B | 2.78 | 3.11 | 3.78 | 3.67 |
| | Average | 2.72 | 2.75 | 3.72 | 3.67 |
| NB-PESQ | Female | 3.81 | 3.82 | 4.09 | 4.06 |
| | Male | 3.66 | 3.68 | 4.07 | 4.08 |
| | Average | 3.75 | 3.76 | 4.06 | 4.07 |

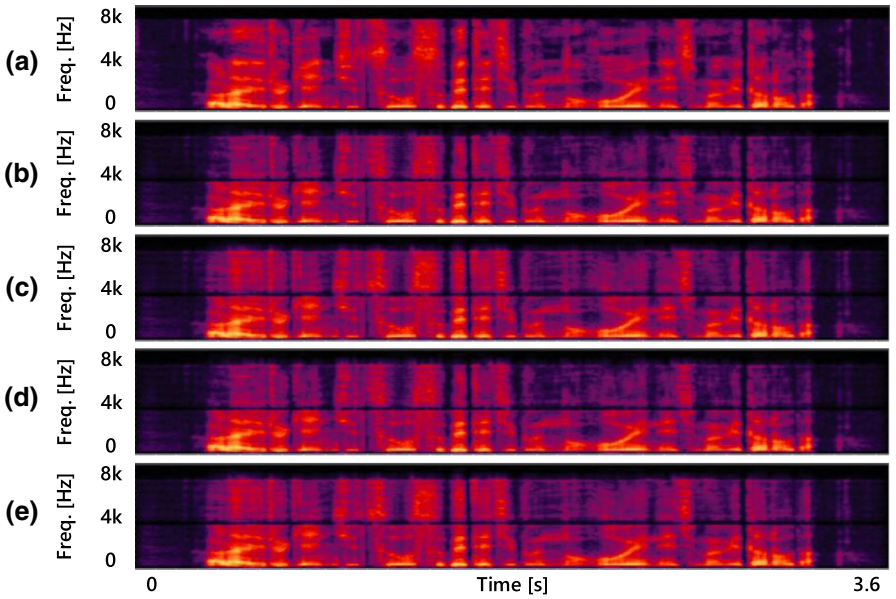


**Fig. 5** Sound spectrograms for original and generated WB speech signals. **a** Original WB speech signal. **b** Generated WB speech signal for BWE-F. **c** Generated WB speech signal for BWE-H. **d** Generated WB speech signal for BWE-FG. **e** Generated WB speech for BWE-HG

FG and BWE-HG) has reconstructed the missing UB spectrum more accurately. Figure 6 shows UB sound pressure changes of the original and generated UB speech signals. It can be seen that the UB sound pressure change of the UB speech signal generated with relative gains was similar to that of the original UB speech signal. We evaluated the distance of the UB sound pressure change between the original and generated UB speech signals by root mean squared error (RMSE). While RMSE for the method of generating a UB speech signal with a relative gain was 12.33 dB, RMSE for the method of generating a UB speech signal with relative gains was 6.18 dB. Therefore, the proposed method achieves the slight UB sound pressure change representation. Table 4 shows results of MOS and LSD. LSD for BWE-HG was lower by 0.15 dB

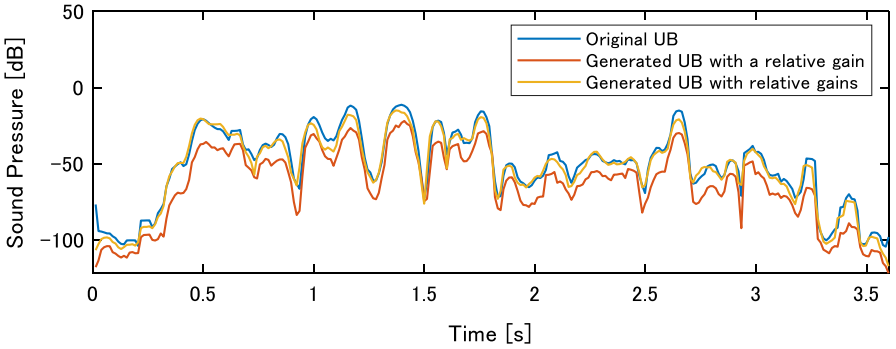


**Fig. 6** UB sound pressure change over time for original UB speech signal (blue line), UB speech signal generated with a relative gain (red line), and UB speech signal with generated relative gains (yellow line)

**Table 4** Subjective and objective quality assessments for generated WB speech signals

| | | BWE-F | BWE-FG | BWE-H | BWE-HG | NB | Original |
|---|---|---|---|---|---|---|---|
| MOS | Female A | 2.40 | 3.20 | 2.80 | 2.80 | 2.60 | 4.80 |
| | Female B | 2.78 | 3.33 | 2.89 | 3.44 | 2.33 | 5.00 |
| | Male A | 2.44 | 3.11 | 2.78 | 3.11 | 1.67 | 4.00 |
| | Male B | 3.00 | 2.89 | 2.89 | 3.22 | 1.67 | 4.33 |
| | Average | 2.69 | 3.16 | 2.87 | 3.15 | 2.03 | 4.58 |
| LSD | Female | 6.32 | 6.20 | 6.32 | 6.20 | – | – |
| | Male | 6.26 | 6.03 | 6.26 | 6.03 | – | – |
| | Average | 6.29 | 6.14 | 6.29 | 6.14 | – | – |

compared to BWE-F. Also, MOS for BWE-HG was over 3.00, which was higher by 1.12 compared to an NB speech signal. These results show that the proposed method enhanced the quality of the NB speech signal more efficiently.

We verified the robustness against artifacts. Table 5 represents the bit error rate of the extracted binary signal under noise environments at several SNR levels. Without speech codecs, a binary signal was extracted successfully from a CNB speech signal in a clean environment. With speech codecs, a bit error occurred even in a clean environment. In particular, G.729 compresses the amount of information in an NB speech signal more than G.711 by quantizing feature vectors, and thus, it was not easy to accurately extract a binary signal from the decoded CNB speech signal. Also, the bit error rate increased as the SNR level decreased. In comparison with the method using TDDH based on the DFT domain (BWE-F and BWE-FG), the method using TDDH based on the DHT domain (BWE-H and BWE-HG) achieved lower bit error rate. These results confirm that the robustness depends on the length of the bandwidth where the hidden vector is embedded, as shown in Eq. (27). While TDDH based on the DFT domain embedded a hidden vector into a magnitude spectrum in the high-frequency bandwidth of 3.4–4 kHz, TDDH based on the DHT domain embedded a hidden vector into an amplitude spectrum in the high-frequency bandwidth of 3.4–4.6

**Table 5** Bit error rate of extracted binary signal under simulation environments at several SNR levels [%]

| | 30 dB | 35 dB | 40 dB | 45 dB | 50 dB | ∞ |
|---|---|---|---|---|---|---|
| *(a)* | | | | | | |
| TDDH based on DFT | 6.97 | 0.87 | 0.00 | 0.00 | 0.00 | 0.00 |
| TDDH based on DHT | 6.76 | 0.81 | 0.00 | 0.00 | 0.00 | 0.00 |
| *(b)* | | | | | | |
| TDDH based on DFT | 8.41 | 2.46 | 1.19 | 1.02 | 0.97 | 0.96 |
| TDDH based on DHT | 8.06 | 2.31 | 1.11 | 0.98 | 0.94 | 0.92 |
| *(c)* | | | | | | |
| TDDH based on DFT | 49.53 | 49.58 | 49.49 | 49.50 | 49.44 | 49.40 |
| TDDH based on DHT | 48.63 | 48.46 | 48.38 | 48.21 | 48.07 | 48.17 |

(a) Without speech codec. (b) G.711. (c) G.729 at 12.2 kbps

kHz. Therefore, the proposed method achieved the robustness improvement against artifacts.

Finally, we discuss the processing time. We have constructed the system with 3.60 GHz Intel i7 core and implemented in MATLAB. Here, the length of the original WB speech signal was 2.22 s. To generate a CNB speech signal, BWE-F, BWE-H, BWE-FG, and BWE-HG took 0.150 s, 0.122 s, 0.151 s, and 0.123 s, respectively. Also, to generate a WB speech signal, BWE-F, BWE-H, BWE-FG, and BWE-HG took 0.120 s, 0.042 s, 0.127 s, and 0.042 s, respectively. The proposed method took longer processing time because relative gains were calculated in some sub-frames and a UB speech signal was generated with 75% overlapping. The total processing time of the proposed method was shorter than the length of the original WB speech signal, and thus, the proposed method worked with less latency communication as well as the conventional method.

## 6 Conclusion

In this paper, we proposed a BWE method using TDDH based on the DHT domain. Subjective listening tests and objective measures showed that the proposed method generated a CNB speech signal without an offset, and thus, the quality difference between NB and CNB speech signals was suppressed. Also, the proposed method generated a UB speech signal with overlapping using relative gains to represent the slight UB sound pressure change over time and enhanced the quality of the NB speech signal by reconstructing the missing UB spectrum. Furthermore, a bit error rate in a noise environment was suppressed by embedding a binary signal into an amplitude spectrum in the wider high-frequency bandwidth. In the future, we will work on speech steganography robust against speech codecs with high compression ratios such as G.729. The code of the proposed method is available at https://github.com/Yuya-Hosoda/Works.

## References


1. J. Abel, T. Fingscheidt, Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 71–83 (2018)
2. Y. Agiomyrgiannakis, Y. Stylianou, Combined estimation/coding of highband spectral envelopes for speech spectrum expansion, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 469–472 (2004)
3. P. Bauer, J. Jones, T. Fingscheidt, Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7039–7042 (2013)
4. S. Chen, H. Leung, Speech bandwidth extension by data hiding and phonetic classification, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 593–596 (2007)
5. S. Chen, H. Leung, A bandwidth extension technique for signal transmission using chaotic data hiding. Circuits Syst. Signal Process. **27**, 893–913 (2008)
6. S. Chen, H. Leung, H. Ding, Telephony speech enhancement by data hiding. IEEE Trans. Instrum. Meas. **56**(1), 63–74 (2007)
7. Z. Chen, C. Zhao, G. Geng, F. Yin, An audio watermark-based speech bandwidth extension method. EURASIP J. Audio Speech Music Process. **10**, 1–8 (2013)
8. N. Cvejic, T. Seppanen, *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks* (IGI Global, Hershey, PA, 2007)
9. E.H. Dinan, B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks. IEEE Commun. Mag. **36**(9), 48–54 (1998)
10. European Telecommunications Standards Institute (ETSI) Standard, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms
11. W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd edn. (Wiley, New York, 1970)
12. B. Geiser, P. Vary, Speech bandwidth extension based on in-band transmission of higher frequencies, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7507–7511 (2013)
13. L. Hanzo, F.C.A. Somerville, J.P. Woodard, *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels* (IEEE Press, Piscataway, NJ, 2001)
14. International Telecommunications Union, Pulse code modulation (PCM) of Voice Frequencies, ITU-T G.711 (1988)
15. International Telecommunications Union, Methods for Subjective Determination of Transmission Quality. ITU-T Recommendation P.800 (1996)
16. International Telecommunications Union, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs
17. International Telecommunications Union, Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB). ITU-T G.722.2 (2003)
18. International Telecommunications Union, Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (SD-ACELP). ITU-T G.729 (2012)
19. K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Jpn. (E) **20**(3), 199–206 (1999)
20. P. Jax, P. Vary, An upper bound on the quality of artificial bandwidth extension of narrowband speech signals, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.237–240 (2002)
21. P. Jax, P. Vary, On artificial bandwidth extension of telephone speech. Signal Process. **83**(8), 1707–1719 (2003)
22. P. Jax, P. Vary, Bandwidth extension of speech signals: a catalyst for the introduction of wideband speech coding? IEEE Commun. Mag. **44**(5), 106–111 (2006)
23. Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design. IEEE Trans. Commun. **COM–28**(1), 84–95 (1980)
24. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)

25. J. Makhoul, M. Berouti, High-frequency regeneration in speech coding systems, in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 428–431 (1979)
26. M. Nilsson, H. Gustaftson, S.V. Andersen, W.B. Kleijn, Gaussian mixture model based mutual information estimation between frequency bands in speech, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 525–528 (2002)
27. M. Nilsson, W. B. Kleijn, Avoiding over-estimation in bandwidth extension of telephony speech, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 869–872 (2001)
28. A. Nishimura, Steganographic band width extension for the AMR codec of low-bit-rate modes, in *Proceeding of Annual Conference of the International Speech Communication Association*, pp. 2611–2614 (2009)
29. N. Prasad, T.K. Kumar, Speech bandwidth extension aided by magnitude spectrum data hiding. Circuits Syst. Signal Process. **36**, 4512–4540 (2017)
30. N. Prasad, G.R.L.V.N.S. Raju, Transform-domain speech bandwidth extension. Circuits Syst. Signal Process. **38**, 5717–5733 (2019)
31. T. Rabie, D. Guerchi, Spectral magnitude speech steganography. Int. J. Comput. Appl. **116**(5), 1–6 (2015)
32. S. Rekik, D. Guerchi, S.A. Selouani, H. Hamam, Speech steganography using wavelet and Fourier transforms. EURASIP J. Audio Speech Music Process. **20**, 1–14 (2012)
33. G. Pirker, M. Wohlmayr, S. Petrik, F. Pernkopf, A pitch tracking corpus with evaluation on multipitch tracking scenario, in *Proceedings of INTERSPEECH*, pp. 1509–1512 (2011)
34. H. Pulakka, U. Remes, S. Yrttiaho, K. Palomaki, M. Kurimo, P. Alku, Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and a gaussian mixture model. IEEE Trans. Audio Speech Lang. Process. **20**(8), 2219–2231 (2012)
35. Y. Qian, P. Kabal, Dual-mode wideband speech recovery from narrowband speech, in *Proceedings of European Conference on Speech Communication and Technology*, pp. 1433–1437 (2003)
36. L.R. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition* (PTR Prentice Hall, Englewood Cliffs, NJ, 1993)
37. A. Sagi, D. Malah, Bandwidth extension of telephone speech aided by data embedding. EURASIP J. Adv. Signal Process. **2007**, 1–16 (2006)
38. T. Unno, A. McCree, A robust narrowband to wideband extension system featuring enhanced codebook mapping, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 805–808 (2005)
39. P. Vary, B. Geiser, Steganographic wideband telephony using narrowband speech codecs, in *Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, pp. 1475–1479 (2007)
40. S. Voran, Listener ratings of speech passbands, in *Proceedings of IEEE Workshop on Speech Coding for Telecommunications*, pp. 81–82 (1997)
41. L. Xiao, X. Dong, The exact transition probability and bit error probability of two-dimensional signaling. IEEE Trans. Wirel. Commun. **4**(5), 2600–2609 (2005)