

Speech Bandwidth Extension Aided by Magnitude Spectrum Data Hiding

N. Prasad¹ · T. Kishore Kumar¹

Received: 5 July 2016 / Revised: 15 February 2017 / Accepted: 16 February 2017 /
Published online: 1 March 2017
© Springer Science+Business Media New York 2017

Abstract Public telephone systems transmit speech across a limited frequency range, about 300–3400 Hz, called narrowband (NB) which results in a significant reduction of quality and intelligibility of speech. This paper proposes a fully backward compatible novel method for bandwidth extension of NB speech. The method uses magnitude spectrum data hiding technique to provide a perceptually better wideband speech signal. The spectral envelope parameters are extracted from the down-sampled frequency shifted version of the high-frequency components of speech signal existing above NB, which are then encoded and spread by using spreading sequences, and are embedded in the low-amplitude high-frequency regions of the magnitude spectrum of NB speech signal. The embedded information is extracted at the receiving end to reconstruct the wideband speech signal. Theoretical and simulation analyses show that the proposed method is robust to quantization and channel noises. The comparison category rating listening and log spectral distortion tests clearly show that the reconstructed wideband signal gives a much better performance in terms of speech quality when compared to the conventional speech bandwidth extension methods employing data hiding.

Keywords Linear prediction · Magnitude spectrum data hiding · Speech bandwidth extension · Speech quality · Spread spectrum · Vector quantization

✉ N. Prasad
Prasad.niz@gmail.com
T. Kishore Kumar
kishorefr@gmail.com

¹ Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Warangal 506004, India

1 Introduction

The human speech contains frequencies beyond the bandwidth of the existing telephone networks which is in the range of 300–3400 Hz. Consequently, the transmission of speech through the telephone networks results in a loss of portions of speech spectrum causing a significant reduction in speech quality and intelligibility. The transmission of wideband (WB) speech which lies in the range of 50–7000 Hz through telephone networks would increase the quality, intelligibility and perceived naturalness of speech signal compared to NB speech. But this requires building of new telephone networks that support larger bandwidths which are expensive and will likely take time to be established [27]. It is, therefore, desirable to enhance the bandwidth at the receiver using speech bandwidth extension (BWE) techniques [24] without modifying infrastructure of the existing telephone networks.

BWE techniques facilitate a major speech quality upgradation of the existing telephone networks. One such method is artificial bandwidth extension (ABE) in which WB signal is reconstructed by estimating the out-of-band (the frequencies below and above NB) information from the NB signal alone. ABE techniques are motivated from the fact that mutual dependence exists between the NB and the out-of-band, which is illuminated from the speech production model. The dependency between the frequency bands justifies the estimation of the out-of-band information from the NB input.

According to the ABE techniques based on source-filter model, BWE is a matter of estimating an excitation signal and a vocal tract filter that modifies the spectral envelope. Estimation of the WB excitation signal is done using noise modulation [34], harmonic and noise modeling [40], sinusoidal synthesis [10], spectral folding [24], spectral translation [24], Pitch adaptive modulation [24] and nonlinear processing [24]. Techniques for estimating the WB spectral envelope include codebooks [36], linear mapping [29], Gaussian mixture models [33], hidden Markov models [2] and neural networks [32]. However, ABE techniques suffer from an inherently limited performance which is not sufficient for the regeneration of high-quality WB speech [25].

A much better WB speech quality compared to ABE techniques is attained when some additional information about the out-of-band is transmitted [27]. In order to ensure the desired backward compatibility with respect to the existing telephone networks, data hiding methods would be used to hide the out-of-band information in the NB signal. Various solutions have already been proposed in the literature for this problem. Siyue Chen and Henry Leung proposed a speech BWE method [3] in which line spectrum pairs of the higher frequencies of speech usually ranging from 3.4 to 7 kHz called as the upperband (UB) signal are encoded and embedded into the NB speech to provide a composite NB speech signal. When the embedded information is extracted and decoded at the receiving end, a perceptually better WB signal is obtained. It was found that this approach gave poor composite NB signal quality. To improve the quality of composite NB signal and reconstructed WB speech [3], Chen and Leung [5] adopted phonetic classification to encode the UB signal more efficiently. But when corrupted by the channel noise and channel spectral distortion, the approaches in [3, 5] gave poor BWE performance. Chen et al. [6] proposed BWE of NB speech by data

hiding in which a hidden channel is provided by the removal of imperceptible components of the NB signal. A good quality WB speech is reconstructed by regenerating the concealed audible components existing outside the NB. The performance of the method is limited by the number of audible components of missing frequency components that can be embedded into the hidden channel. Chen et al. [7] proposed BWE of telephony speech by perception based least significant bit watermark method which embeds UB components into the NB speech to be extracted at the receiver for reconstructing a high-quality WB signal. Vary and Geiser [39] used joint coding and data hiding to embed 600 bit/s of additional information into the NB codec bit stream to develop a backward compatible WB codec. A backward compatible WB telephony based on NB coder and on BWE with 400 bit/s of additional information which is embedded into the NB codec bit stream was proposed by Geiser and Vary [15]. This approach gave poor performance when codec bits are corrupted by channel noise.

Speech-in-speech hiding used for reducing the storage and transmission requirements of electronic voice mail, proposed in [35], employs magnitude spectrum data hiding to embed the parameters of secret speech signal in the low-amplitude high-frequency regions of the spectral magnitude of the host speech signal without degrading host signal perceptual quality. It was found that this approach is producing a stego speech signal that is indistinguishable from the host speech, while being able to extract the secret speech signal without any degradation in quality.

In digital telephone network, only the speech BWE toward UB is considered. This is because the transmission at the low-frequency end (0–300 Hz) is usually not a problem in a digital telephone network [9, 17, 23]. In this paper, a novel NB speech BWE technique using magnitude spectrum data hiding [35] for extending the bandwidth of the digital telephone network is proposed. That is, the down-sampled frequency shifted version of UB signal is analyzed using linear predictive coding (LPC), and the extracted spectral envelope parameters are then encoded and embedded in the low-amplitude high-frequency part of the magnitude spectrum of NB speech without altering the low-frequency part. At the receiver, when the embedded information is extracted, a WB signal with a much better speech quality can be reconstructed by combining the UB signal that was transmitted through the low-amplitude high-frequency part of the magnitude spectrum and the NB signal. The proposed scheme uses the real UB information instead of its estimation which makes the reconstruction of the WB speech more accurate compared to the conventional ABE methods. Also, the proposed method is compatible with conventional NB terminal equipments, e.g., a plain ordinary telephone set (POTS). In other words, conventional NB receivers can still access the NB speech properly without additional hardware, while a customized receiver is able to extract the embedded information and provide WB signal with much better quality.

Along with quantization noise (QN) and channel noise, the telephone network channel introduces spectral distortion. Techniques proposed in [3, 5, 14] for BWE of NB speech treats only the QN [pulse code modulation (PCM), μ -law, adaptive delta pulse code modulation (ADPCM), enhanced full rate (EFR)] ignoring the channel noise and spectral distortion. In this work, direct sequence code division multiple access (DS-CDMA) technique is employed to suppress channel noise and QN, and adaptive channel equalization is employed to compensate the effect of spectral distortion.

DS-CDMA technique is employed in this work for the successful extraction of the embedded information as it is well recognized by its robustness to interference. In particular, specific spreading sequence (SS) is multiplied with each data bit to be embedded to spread it. The spread signals are then summed up together to form the embedded information. Since the spreading sequences (SSs) being orthogonal over one another and have a low cross-correlation, the embedded information can be reliably recovered by using a multiuser detector [31].

To minimize the interference caused by the other embedded components, spreading sequences with low cross-correlations are preferred. Hadamard codes have an optimal cross-correlation performance, i.e., orthogonal to each other, whereas the m -sequences, Gold-codes and Kasami-codes are with varying cross-correlation properties [8, 16]. Because the Hadamard codes are well recognized by its optimal cross-correlation performance, it is employed in this work for minimizing the interference caused by the other embedded components.

Adaptive channel equalization [37] is also employed in this paper to mitigate the degradation of performance due to spectral distortion caused by a telephone network channel. Since recursive least square (RLS) algorithm converges fast and has good tracking capability, it is used for channel equalization.

The rest of the paper is organized as follows. In Sect. 2, we introduce the magnitude spectrum data hiding based approach for BWE. Section 3 describes the proposed NB speech BWE method using magnitude spectrum data hiding technique in detail. The performances of the proposed method under additive white Gaussian noise (AWGN) and QN are analyzed in Sect. 4. Section 5 discusses the subjective and objective test results of the proposed method. Finally, conclusions are presented in Sect. 6.

2 Magnitude Spectrum Data Hiding Scheme

2.1 Fourier Transform of NB Speech

The discrete Fourier transform (DFT) of a NB signal, $X_{\text{NB}}(n)$, $0 \leq n \leq N - 1$, is given by

$$X_{\text{NB}}(k) = \sum_{n=0}^{N-1} X_{\text{NB}}(n) e^{-j2\pi kn/N}, \quad 0 \leq k \leq m - 1 \quad (1)$$

In polar coordinates, $X_{\text{NB}}(k)$, can be expressed as:

$$X_{\text{NB}}(k) = |X_{\text{NB}}(k)| e^{j\Phi_{\text{NB}}(k)} \quad (2)$$

The magnitude spectrum $|X_{\text{NB}}(k)|$ satisfies the symmetry shown in Eq. (3) and hence can be defined uniquely by the first $m/2$ frequency components.

$$|X_{\text{NB}}(k)| = |X_{\text{NB}}(m - k)| \quad (3)$$

This symmetry must be preserved when embedding information in the magnitudes.

2.2 Properties of NB Speech

The distribution of the first three formants of NB speech denotes the primary cues to the vowels and most of the vowel energy is concentrated below 1000 Hz [38]. The spectral magnitude shows very weak components at high frequencies [38] for all vowels and most of the voiced consonants. Few unvoiced fricative consonants present large magnitudes at high frequencies, and even if we do not model correctly these high-frequency components [1] the NB signal intelligibility is negligibly affected. This has motivated us to hide the spreaded spectral envelope parameters of the down-sampled frequency shifted version of UB signal in the low-amplitude high-frequency region of the magnitude spectrum of NB signal.

2.3 Magnitude Spectrum Data Hiding for Bandwidth Extension

In the first stage of the magnitude spectrum data hiding algorithm, the fast Fourier transform (FFT) is performed on NB signal $X_{\text{NB}}(n)$, followed by decomposition into its magnitude spectrum $|X_{\text{NB}}(k)|$ ($k = 0, 1, \dots, m - 1$) and phase spectrum $\Phi_{\text{NB}}(k)$. We assume that the down-sampled frequency shifted version of UB signal is encoded into a sequence of data bits, i.e., $\{d_r\}$, $d_r \in \{-1, 1\}$, $r = 0, 1, \dots, R - 1$, where R denotes the total number of bits.

Each data bit to be embedded have to be spread by multiplying with a specific SS (Hadamard codes are used in this paper), i.e.,

$$d_r c^r, \quad 0 \leq r \leq R - 1 \quad (4)$$

where c^r is a SS with length U , and r is the index of a particular data bit (d_r) and its corresponding SS used to spread it. The spreading vectors are then summed up together to form the embedded information given by

$$H = \sum_{r=0}^{R-1} d_r c^r \quad (5)$$

where H denotes the vector of embedded information. The U last elements of the first half of the $|X_{\text{NB}}(k)|$ are replaced by βH of the down-sampled frequency shifted version of UB signal [35]. The resulting magnitude spectrum, denoted by $|X^1_{\text{NB}}(k)|$, is given by (6)

$$|X^1_{\text{NB}}(k)| = \begin{cases} |X_{\text{NB}}(k)|, & k = 0, \dots, \frac{m}{2} - U \\ \beta H_u, & k = \frac{m}{2} - U - 1, \dots, \frac{m}{2} - 1 \\ \beta H_u, & k = \frac{m}{2}, \dots, \frac{m}{2} + U \\ |X_{\text{NB}}(k)|, & k = \frac{m}{2} + U + 1, \dots, m - 1 \end{cases} \quad (6)$$

where H_u is the u th element of H , and β is a scalar that will improve the composite NB signal quality X^1_{NB} , i.e.,

$$\beta^2 H_u^2 \leq \frac{1}{E_{\text{NB}}} \quad (7)$$

where E_{NB} is the energy of $|X_{\text{NB}}(k)|$. Hence, a suitable value of β can be determined by $\beta = \sqrt{\frac{1}{E_{\text{NB}} H_u^2}}$. Considering that $|d_r c_u^r|$ is always equal to 1, we have

$$\beta = \sqrt{\frac{1}{R E_{\text{NB}}}} \quad (8)$$

These modifications lead to a composite NB signal and its spectrum is a combination of $|X^1_{\text{NB}}(k)|$ and $\Phi_{\text{NB}}(k)$,

$$X^1_{\text{NB}}(k) = |X^1_{\text{NB}}(k)| e^{j\Phi_{\text{NB}}(k)}, \quad k = 0, \dots, m-1 \quad (9)$$

The time domain composite NB speech signal $X^1_{\text{NB}}(n)$, $n = 0, \dots, N-1$, is obtained by the inverse FFT (IFFT) and it is transmitted through the telephone network channel.

$$X^1_{\text{NB}}(n) = \text{IFFT}(X^1_{\text{NB}}(k)) \quad (10)$$

The composite NB signal $X^1_{\text{NB}}(n)$, communicated through the telephone network channel is distorted and noisy. If the proposed speech BWE receiver operate on the distorted signal it would give a large bit error rate (BER). To compensate the channel spectral distortion, various channel equalization methods have already been proposed in the literature [37]. An equalizer that performs better, in terms of a lower mean square error (MSE), will usually result in a lower BER in embedded information extracting. Hence in this paper, the RLS algorithm with 512 taps was employed for channel equalization. Let $\hat{X}^1_{\text{NB}}(n)$ denote the signal observed at the output of channel equalizer. $\hat{X}^1_{\text{NB}}(n)$ can be assumed to be perceptually similar to $X^1_{\text{NB}}(n)$ since the channel is able to communicate $X^1_{\text{NB}}(n)$ with an adequate perceptual quality. Thus, we have

$$X^1_{\text{NB}}(n) \approx \hat{X}^1_{\text{NB}}(n) \quad (11)$$

Furthermore, the perceived differences between $X_{\text{NB}}(n)$ and $X^1_{\text{NB}}(n)$ would be very small since only the low-amplitude high-frequency part of $|X_{\text{NB}}(k)|$ were changed to obtain $X^1_{\text{NB}}(n)$. Hence, $X^1_{\text{NB}}(n)$ can be approximated by $X_{\text{NB}}(n)$

$$X^1_{\text{NB}}(n) \approx X_{\text{NB}}(n) \quad (12)$$

i.e., the quality of $X_{\text{NB}}(n)$ is not considerably degraded although additional information is embedded into it. From Eqs. (11) and (12), we have

$$\hat{X}^1_{\text{NB}}(n) \approx X_{\text{NB}}(n) \quad (13)$$

i.e., the perceptual quality of the received $\hat{X}_{NB}^1(n)$ is similar to that of the NB signal $X_{NB}(n)$.

In order to retrieve the down-sampled frequency shifted version of UB signal, initially FFT is applied to $\hat{X}_{NB}^1(n)$ and decompose into its magnitude spectrum and phase spectrum. The embedded information is then extracted from the same locations as they were embedded in the magnitude spectrum of $\hat{X}_{NB}^1(n)$ [35] by

$$\hat{H}_u = \left| \hat{X}_{NB}^1(k) \right|, \quad k = \frac{m}{2} - U - 1, \dots, \frac{m}{2} - 1 \tag{14}$$

We use \hat{H}_u instead of βH_u to denote the extracted embedded information because it is subjected to noise corruption. A multiuser detector is employed here to decode the data bits. That is,

$$\hat{d}_r = \text{sign} \left(\sum_{u=0}^{U-1} \hat{H}_u c_u^r \right) \tag{15}$$

In a noise-free environment, $\hat{H}_u = \beta H_u$. Substituting it into (15), the decoding process can be further interpreted as

$$\begin{aligned} \hat{d}_r &= \text{sign} \left(\sum_{u=0}^{U-1} \beta H_u c_u^r \right) \\ &= \text{sign} \left(\beta \sum_{u=0}^{U-1} \left(d_r c_u^r c_u^r + \sum_{q=0, q \neq r}^{R-1} d_q c_u^q c_u^r \right) \right) \\ &= \text{sign} \left(\beta U d_r + \beta \sum_{q=0, q \neq r}^{R-1} d_q \sum_{u=0}^{U-1} c_u^q c_u^r \right) \end{aligned} \tag{16}$$

The employed SSs are mutually orthogonal. That is

$$\sum_{u=0}^{U-1} c_u^q c_u^r = 0 \tag{17}$$

where $q \neq r$. Therefore,

$$\beta \sum_{q=0, q \neq r}^{R-1} d_q \sum_{u=0}^{U-1} c_u^q c_u^r = 0 \tag{18}$$

It can be seen that if $\beta \sum_{q=0, q \neq r}^{R-1} d_q \sum_{u=0}^{U-1} c_u^q c_u^r = 0$, the interference caused by the other embedded components can be completely removed. This shows that the data bits, i.e., d_r , that represent the down-sampled frequency shifted version of UB signal can be successfully recovered due to the suppression of the interference caused by the other embedded components because of employing DS-CDMA technique.

3 Bandwidth Extension of NB Speech Using Magnitude Spectrum Data Hiding

In this section, we propose to extend the bandwidth of NB speech by the proposed magnitude spectrum data hiding technique. The basic idea is the spreaded spectral envelope parameters of the down-sampled frequency shifted version of UB signal are embedded in the low-amplitude high-frequency part of $|X_{NB}(k)|$ at the transmitter. The embedded information is extracted at the receiving end to reconstruct the WB speech signal with much better quality. The transmitter and receiver sections of the proposed NB speech BWE method using magnitude spectrum data hiding technique are discussed below.

3.1 Transmitter

The proposed speech BWE transmitter is shown in Fig. 1. Initially, the original WB speech (0–7 kHz), with a sampling rate of 16 kHz, is band-splitting using a low pass filter (LPF) and a high pass filter (HPF), respectively. The LPF output (0–3.5 kHz) is then decimated to provide the NB signal, denoted by $X_{NB}(n)$. The HPF output (3.5–7 kHz) is shifted to the NB frequency range, and also decimated to provide the extendedband signal, denoted by $X_{EB}(n)$. Thus, the sampling rate of $X_{NB}(n)$ and $X_{EB}(n)$ is 8 kHz, which is the sampling rate of the channel.

To embed $X_{EB}(n)$ into the NB signal imperceptibly, it would be required to minimize the number of the data bits, i.e., d_r , that represent $X_{EB}(n)$. Here, the LPC [18] and vector quantization (VQ) are employed to accomplish this target. LPC is based on the source-filter model of speech production. The filter coefficients are the reciprocal of the autoregressive (AR) filter coefficients. The AR coefficients which correspond to the spectral envelope of $X_{EB}(n)$ are denoted as b_i ($i = 1, \dots, 10$), where i is the order of filter, which are found by using the Levinson–Durbin algorithm [18] to solve the set of equations

$$\sum_{i=1}^{10} b_i a^1(|i - q|) = -a^1(q), \quad q = 1, \dots, 10, \quad (19)$$

where $a^1(q)$ are the modified autocorrelation coefficients. These AR coefficients are then converted to line spectral frequencies (LSFs). This is because slight change in AR coefficients will result in significant distortions when reconstructing $X_{EB}(n)$. The extendedband signal $X_{EB}(n)$ gain should also be embedded since the reconstructed $X_{EB}(n)$ has to be scaled to an appropriate energy to circumvent over-estimation [30]. Therefore, the relative gain of $X_{EB}(n)$ against $X_{NB}(n)$ is calculated as $GA_{rel} = \frac{GA_{EB}}{GA_{NB}}$ and combined with 10 LSFs to provide a representation vector of $X_{EB}(n)$, i.e., $D = [LSF_1, LSF_2, \dots, LSF_{10}, GA_{rel}]$.

We do not encode the excitation signal parameters of $X_{EB}(n)$ for minimizing the number of the data bits to be embedded. This is because, human ear is insensitive to the excitation signal distortions at frequencies above 3400 Hz [26]. Therefore, $X_{EB}(n)$

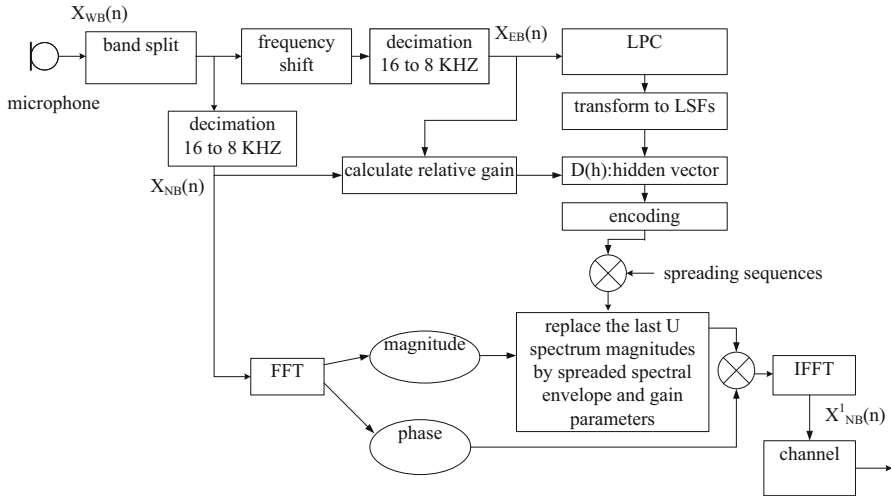


Fig. 1 Proposed speech BWE transmitter

excitation estimation from the NB signal at the receiver is well-suited for the reconstruction performance.

The representation vector D is then quantized to the closest entry of a VQ codebook that is generated by the Linde–Buzo–Gray training (LBG) algorithm [28]. The binary representation of the entry index, i.e., $(d_0d_1d_2 \dots d_{R-1})$ is embedded into NB signal by the proposed magnitude spectrum data hiding technique to produce a composite signal $X_{NB}^1(n)$ that can be transmitted through the telephone network channel.

To achieve synchronization [11] among the transmitting and receiving frames, a synchronization sequence such as 111...11 is inserted at each frame of the composite NB signal. Reception of a certain number of successive identical waveforms (synchronization sequence) at the receiver indicates the arrival of a new composite NB signal frame.

3.2 Receiver

The proposed speech BWE receiver is shown in Fig. 2. At the receiver, frame synchronization is first performed. When the embedded information is extracted by the proposed magnitude spectrum data hiding technique, the entry index can be obtained and the corresponding quantized LSFs and relative gain are properly retrieved from the VQ codebook. The LSFs are transformed back to the AR coefficients. Meanwhile, the excitation signal is obtained as the residue of an LPC analysis on $\hat{X}_{NB}^1(n)$, i.e.,

$$res(n) = \hat{X}_{NB}^1(n) - \sum_{j=1}^{10} c_j \hat{X}_{NB}^1(n - j) \tag{20}$$

where c_j denotes the AR coefficients of $\hat{X}_{NB}^1(n)$ and $res(n)$ is the residue (excitation). It is then fed into the synthesis filter described by the AR coefficients to reconstruct

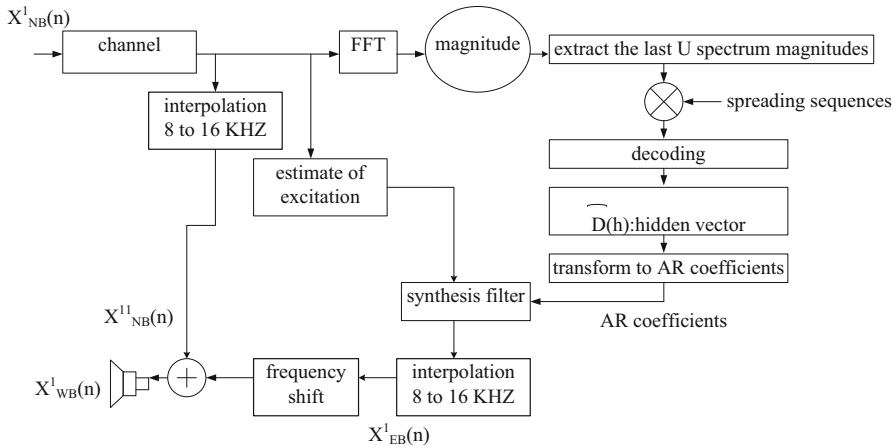


Fig. 2 Proposed speech BWE receiver

the extendedband signal $\hat{X}_{EB}(n)$ that was embedded. The $\hat{X}_{EB}(n)$ gain is adjusted to $G\hat{A}_{EB}$, which is obtained by $G\hat{A}_{EB} = G\hat{A}_{rel} \cdot G\hat{A}_{NB}$, where $G\hat{A}_{NB}$ is estimated from $\hat{X}_{NB}^1(n)$, $G\hat{A}_{rel}$ is obtained from the \hat{D} . At this point, the sample rate for both the $\hat{X}_{NB}^1(n)$ and $\hat{X}_{EB}(n)$ is 8000Hz and should be interpolated to 16,000Hz, the sample rate for WB speech. The interpolated $\hat{X}_{EB}(n)$, denoted by $X_{EB}^1(n)$, lies in the NB frequency range and is shifted to UB. Finally, perceptually better WB speech ($X_{WB}^1(n)$) is reconstructed by adding the restored extendedband ($X_{EB}^1(n)$) signal to the up-sampled composite NB ($X_{NB}^{11}(n)$).

4 Performance Analysis of Magnitude Spectrum Data Hiding

Although using orthogonal SSs can successfully recover d_r in a noise-free transmission, in practice, there is always noise corruption during transmission. The most common one is AWGN. Let $\hat{x}_{NB}^1(n)$ denote the received NB signal observed at the channel equalizer output and is given by (21)

$$\hat{x}_{NB}^1(n) = x_{NB}^1(n) + g(n) \tag{21}$$

where $g(n)$ is AWGN with zero mean and a variance of σ_g^2 . Its FFT $\hat{x}_{NB}^1(k)$ is given by

$$\hat{x}_{NB}^1(k) = \sum_{n=0}^{N-1} \hat{x}_{NB}^1(n) e^{-j \frac{2\pi kn}{N}} = \sum_{n=0}^{N-1} (x_{NB}^1(n) + g(n)) e^{-j \frac{2\pi kn}{N}} \tag{22}$$

In other words, the extracted embedded information is given by

$$\hat{H}_u = \beta H_u + \sum_{n=0}^{N-1} g(n) e^{-j \frac{2\pi kn}{N}} \tag{23}$$

based on (6) and (14). Defining $J_u = \sum_{n=0}^{N-1} g(n)e^{-j\frac{2\pi un}{N}}$, (23) becomes

$$\hat{H}_u = \beta H_u + J_u \tag{24}$$

A multiuser detector is employed here to decode the data bits. That is

$$\begin{aligned} \hat{d}_r &= \text{sign} \left(\sum_{u=0}^{U-1} (\beta H_u + J_u) c_u^r \right) \\ &= \beta U d_r + \sum_{q \neq r}^{R-1} \beta d_q \sum_{u=0}^{U-1} c_u^q c_u^r + \sum_{u=0}^{U-1} J_u c_u^r \end{aligned} \tag{25}$$

Since the interference caused by the other embedded components can be completely removed because of orthogonality of SSs, (25) can be further written as

$$\hat{d}_r = \beta U d_r + \sum_{u=0}^{U-1} J_u c_u^r \tag{26}$$

From (26), a detection error occurs if $\left| \sum_{u=0}^{U-1} J_u c_u^r \right| \geq \beta U$ and $\sum_{u=0}^{U-1} J_u c_u^r$ has an opposite sign with d_r . According to the central limit theorem [12], the conditional probability density function (CPDF) of \hat{d}_r is given by

$$p(\hat{d}_r / -1) = \frac{1}{\sqrt{2\pi\sigma_{Jc}^2}} e^{-\frac{(\hat{d}_r + \beta U)^2}{2\sigma_{Jc}^2}} \tag{27}$$

provided $d_r = -1$. Here, σ_{Jc}^2 is the variance of the random variable $\sum_{u=0}^{U-1} J_u c_u^r$, which can be expressed by

$$\sigma_{Jc}^2 = E \left[\left(\sum_{u=0}^{U-1} J_u c_u^r \right)^2 \right] = E \left[\sum_{u=0}^{U-1} \sum_{k=0}^{U-1} J_u J_k c_u^r c_k^r \right] \tag{28}$$

Considering that c_u^r is a random variable with zero mean and a variance of 1, and J_u is uncorrelated with c_u^r , (28) can be further written as

$$\sigma_{Jc}^2 = E \left[\sum_{u=0}^{U-1} J_u^2 (c_u^r)^2 \right] = \sum_{u=0}^{U-1} E \left[J_u^2 (c_u^r)^2 \right] = U \sigma_J^2 \tag{29}$$

where σ_J^2 denotes the variance of J_u . Analogous to (27), the CPDF for $d_r = 1$ can be formulated as

$$p(\hat{d}_r / 1) = \frac{1}{\sqrt{2\pi\sigma_{Jc}^2}} e^{-\frac{(\hat{d}_r - \beta U)^2}{2\sigma_{Jc}^2}} \tag{30}$$

Therefore, the CPDF of deciding in favor of $\hat{d}_r = 1$ when $d_r = -1$. That is,

$$p(1/-1) = \int_0^\infty p(\hat{d}_r/-1)d\hat{d}_r = \frac{1}{\sqrt{2\Pi\sigma_{Jc}^2}} \int_0^\infty e^{-\frac{(\hat{d}_r+\beta U)^2}{2\sigma_{Jc}^2}} d\hat{d}_r \tag{31}$$

The above equation can be expressed in terms of the complementary error function (CEF) as

$$p(1/-1) = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{\beta^2 U^2}{2\sigma_{Jc}^2}} \right) \tag{32}$$

where $\operatorname{erfc}(q) = \frac{2}{\sqrt{\Pi}} \int_q^\infty e^{-t^2} dt$. Similarly, we can derive the CEF for deciding in favor of $\hat{d}_r = -1$ when $d_r = 1$. That is,

$$p(-1/1) = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{\beta^2 U^2}{2\sigma_{Jc}^2}} \right) \tag{33}$$

Based on $p(1/-1)$ and $p(-1/1)$, the average probability of detection error is given by

$$p_{\text{error}} = p(-1)p(1/-1) + p(1)p(-1/1) = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{\beta^2 U^2}{2\sigma_{Jc}^2}} \right) \tag{34}$$

assuming that $d_r = -1$ and $d_r = 1$ are equiprobable. Substituting (29) and (8) into (34), we have

$$p_{\text{error}} = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{U}{2RE_{\text{NB}}\sigma_J^2}} \right) \tag{35}$$

Recalling that $J_u = \sum_{n=0}^{N-1} g(n)e^{-j\frac{2\Pi un}{N}}$, σ_J^2 can be obtained as

$$\sigma_J^2 = E \left[\left(\sum_{n=0}^{N-1} g(n)e^{-j\frac{2\Pi un}{N}} \right)^2 \right] = E \left[\sum_{n=0}^{N-1} g^2(n)e^{-j\frac{4\Pi un}{N}} \right] = \sigma_g^2 \sum_{n=0}^{N-1} e^{-j\frac{4\Pi un}{N}} \tag{36}$$

Since $\sum_{n=0}^{N-1} e^{-j\frac{4\Pi un}{N}}$ can be treated as a scalar of σ_g^2 when N is set as a constant, we can reduce (36) to

$$\sigma_J^2 = B\sigma_g^2 \tag{37}$$

where $B = \sum_{n=0}^{N-1} e^{-j\frac{4\Pi un}{N}}$. Substituting (37) into (35), the error probability under AWGN can be obtained as

$$p_{\text{error}} = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{U}{2RE_{\text{NB}}B\sigma_g^2}} \right) \tag{38}$$

The performance of the proposed scheme under AWGN depends on the length of SS, the number of data bits that represent $X_{\text{EB}}(n)$, the energy of NB signal and variance of AWGN.

Besides AWGN, the embedded information is also required to be robust to the QN. The quantized NB signal is given by

$$\hat{x}_{\text{NB}}^1(n) = x_{\text{NB}}^1(n) + g'(n) \tag{39}$$

where $g'(n)$ is the QN. Its FFT $\hat{x}_{\text{NB}}^1(k)$ is given by

$$\hat{x}_{\text{NB}}^1(k) = \sum_{n=0}^{N-1} \hat{x}_{\text{NB}}^1(n) e^{-j \frac{2\pi kn}{N}} = \sum_{n=0}^{N-1} (x_{\text{NB}}^1(n) + g'(n)) e^{-j \frac{2\pi kn}{N}} \tag{40}$$

Correspondingly, the extracted embedded information is given by

$$\hat{H}_u = \beta H_u + \sum_{n=0}^{N-1} g'(n) e^{-j \frac{2\pi kn}{N}} \tag{41}$$

Comparing (41) with (23), except that $g'(n)$ is used instead of $g(n)$, the extracted embedded information expressions under AWGN and under QN have the same formula. Hence, if we define $J_u = \sum_{n=0}^{N-1} g'(n) e^{-j \frac{2\pi kn}{N}}$, the derivation of the error probability under QN should be the similar as that under AWGN. In other words, the error probability under QN is given by

$$p_{\text{error}} = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{U}{2RE_{\text{NB}}\sigma_J^2}} \right) \tag{42}$$

When a uniform quantization (UQ) is applied, $X_{\text{NB}}(n)$ can be modeled as being uniformly distributed within each cell, i.e., $[l\Delta - \frac{\Delta}{2}, l\Delta + \frac{\Delta}{2}]$, where l is an integer and Δ is the step size. Thus, we can obtain the expected squared error by UQ as

$$\sigma_g^2 = E \left[g^2(l) \right] = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} g^2(l) p(g'(l)) dg'(l) = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta} g^2(l) dg'(l) = \frac{\Delta^2}{12} \tag{43}$$

Substituting (43) into (37), we have

$$\sigma_J^2 = B \frac{\Delta^2}{12} \tag{44}$$

Given (42) and (44), the error probability under UQ is given by

$$p_{\text{error}} = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{6U}{\operatorname{BRE}_{\text{NB}} \Delta^2}} \right) \quad (45)$$

Besides UQ, non-UQ is also popular in analog-to-digital conversion of speech signals. In non-UQ, the signal to quantization error ratio (SQER) is required to lie within 38 dB [18], so that the digitized speech quality can be retained. That is, assuming the power of NB speech as σ_X^2 , we have $\frac{\sigma_X^2}{\sigma_s^2} = 38 \text{ dB} = 6309.6$. Hence, the error probability under non-UQ is given by

$$p_{\text{error}} = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{U}{2 \times 1.585 \times 10^{-4} \operatorname{BRE}_{\text{NB}} \sigma_X^2}} \right) = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{3154.8U}{\operatorname{BRE}_{\text{NB}} \sigma_X^2}} \right) \quad (46)$$

The performance of the proposed scheme under non-UQ depends on the length of SS, the number of data bits that represent $X_{\text{EB}}(n)$, the energy and power of NB speech.

5 Experimental Results

The speech samples used for evaluation of the proposed method were taken from the TIMIT database [13]. Ten different sentences spoken by ten different male and female speakers of 2–2.5 s long were taken for evaluating the performance of the proposed method. ANB reference samples were produced by the adaptive multirate (AMR) NB coded NB samples. AWB references were produced by the AMR-WB coded WB samples. The NB samples were partitioned into non-overlapping 20 ms frames to be processed one by one.

The performance evaluation was done using both subjective and objective measures. The different methods compared with the proposed method are ABE of telephony speech by data hiding [3], speech BWE by data hiding and phonetic classification [5], telephony speech enhancement by data hiding [6], An audio watermark-based speech BWE [7], steganographic WB telephony using narrowband speech codecs [39] and ABE of speech supported by watermark transmitted side information [14] represented, respectively, by conventional data hiding (CDH), data hiding with phonetic classification (DHWPC), conventional signal domain data hiding (SDDH), conventional bit stream data hiding (BSDH), conventional joint coding and data hiding (JCDH) and conventional WTSI (CWTSI) in the analysis. Conventional WTSI uses the vectorial form of quantization modulation index (QIM) for speech BWE. The experiments conducted in this paper use the two channel models provided below:

- (i) μ -law channel model.
- (ii) AWGN channel model with a signal to noise ratio (SNR) of 35 dB.

Table 1 Comparison mean opinion score (CMOS)

Score	Rating of second signal compared to that of the first signal
3	Much better
2	Better
1	Slightly better
0	About the same
−1	Slightly worse
−2	Worse
−3	Much worse

5.1 Subjective Quality Evaluation

The obtained speech quality of the proposed method in comparison with the ANB reference sample, AWB reference sample and outputs of the conventional speech BWE methods [3, 5–7, 14, 39] was assessed using comparison category rating (CCR) listening test [19] recommended by international telecommunications union (ITU-T) and perceptual transparency was assessed using mean opinion scores (MOS) test [3–5]. A subjective test is employed to evaluate the comparative performance of the bandwidth extended NB telephonic speech taken from NTIMIT database with the corresponding WB speech signal taken from TIMIT database. The subjective comparison of original WB speech taken from the TIMIT database, composite NB speech, telephone speech taken from the NTIMIT database and reconstructed WB speech by simulating NB quality speech from original WB speech signal using ITU tools was also employed. In these tests, the speech samples have been presented to each listener through headphones, separately in a quiet room and the evaluation was made by asking people's opinions on speech sounds using a predefined scale. Twenty normal listeners (10 females and 10 males) between the age of 22 and 32 years participated in these tests.

5.1.1 Pairwise Comparisons

CCR listening test compares two speech samples and provides information on which sample is better in terms of quality based on the comparison mean opinion score (CMOS) given in Table 1. Subjects participating in a CCR listening test compared pairs of speech samples from the ANB, AWB, outputs of the conventional speech BWE methods [3, 5–7, 14, 39] and proposed method $X^1_{WB}(n)$. The second signal of the pair is rated compared to the first signal. The average listener ratings in pairwise comparisons between the ANB, AWB, outputs of the conventional speech BWE methods [3, 5–7, 14, 39] and $X^1_{WB}(n)$ are shown in Fig. 3a, b in which the relative frequencies of the scores are indicated by the bars. Bars at positive score values indicate the preference for the second signal in the pairwise comparison. For example, in the pairwise comparison between the ANB and the AWB, bars on positive side show preference for the AWB.

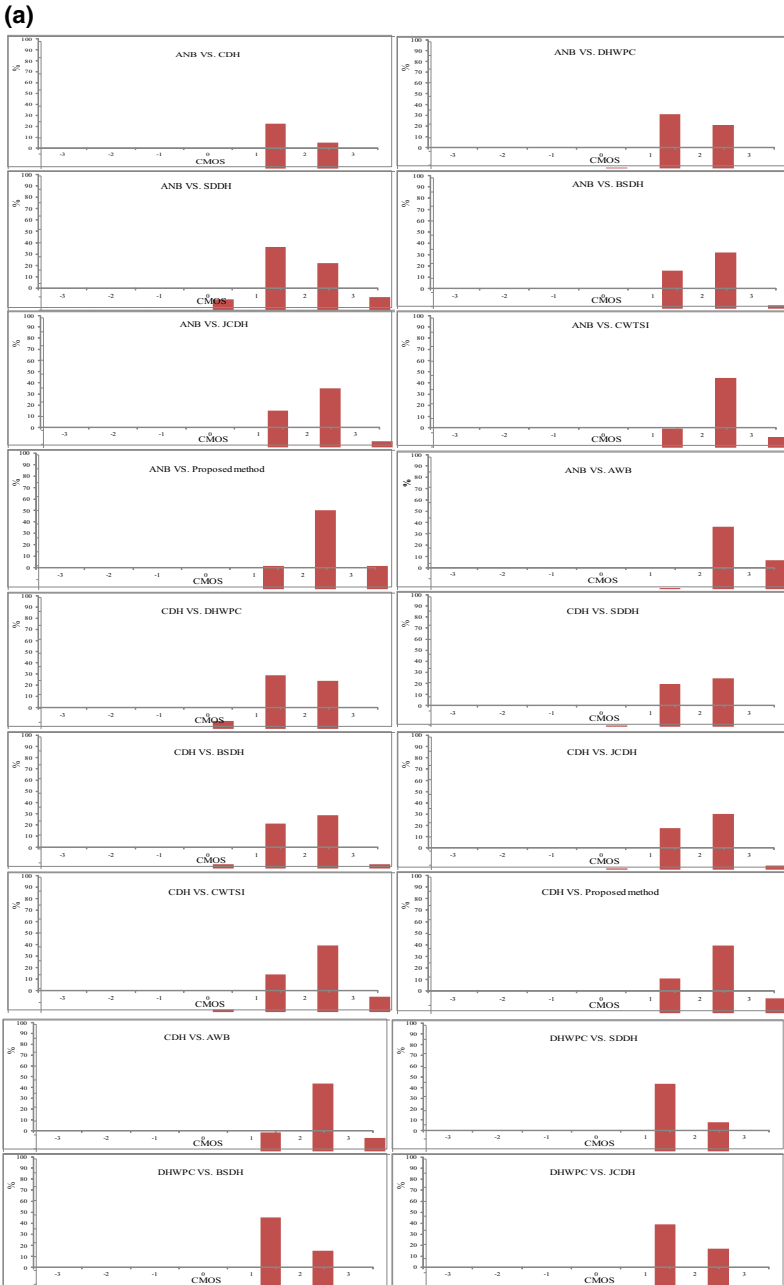


Fig. 3 a Results of pairwise comparisons of ANB, AWB, CDH [3], DHWPC [5], SDDH [6], BSDH [7], JCDH [39], CWTSI [14] and proposed method. In each illustration, the bars indicate relative frequencies of the scores from much worse (-3) to much better (3) in pairwise comparisons. **b** Results of pairwise comparisons of ANB, AWB, CDH [3], DHWPC [5], SDDH [6], BSDH [7], JCDH [39], CWTSI [14] and proposed method. In each illustration, the bars indicate relative frequencies of the scores from much worse (-3) to much better (3) in pairwise comparisons

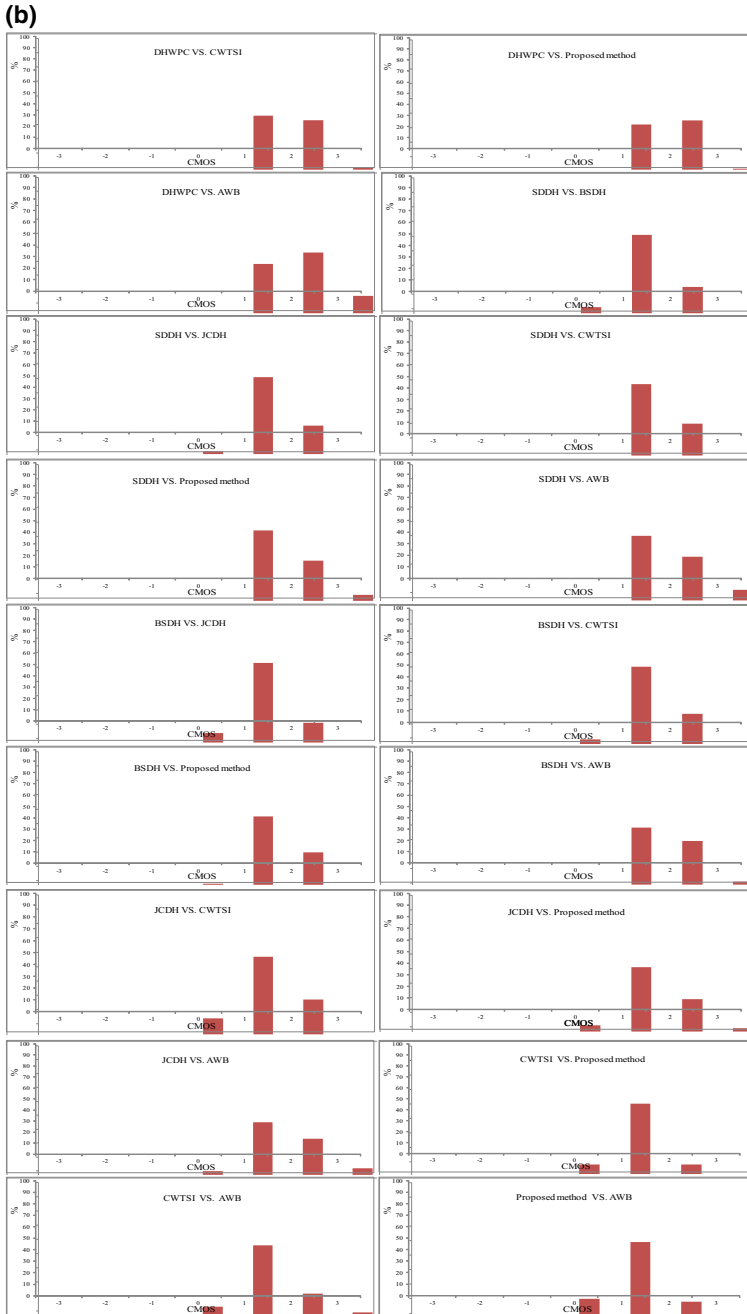


Fig. 3 continued

The distributions in Fig. 3a, b indicate the superiority of the AWB over the ANB, outputs of the conventional speech BWE methods [3, 5–7, 14, 39] and $X^1_{WB}(n)$, and that of $X^1_{WB}(n)$ over the ANB and outputs of the conventional speech BWE methods [3, 5–7, 14, 39] in terms of speech quality.

5.1.2 Perceptual Transparency

The proposed method should embed information transparently. That is, the composite NB signal $X^1_{NB}(n)$ should be subjectively indistinguishable from NB signal $X_{NB}(n)$. MOS test [3–5] is used in this paper to assess the perceptual transparency. Subjects participating in an MOS test compare pairs of speech samples from $X_{NB}(n)$ and $X^1_{NB}(n)$, and give their opinions in terms of MOS given in Table 2. Table 3 presents the resultant average MOS over all subjects and all samples of conventional speech BWE methods [3, 5–7, 14, 39] and proposed method. A clear perceptual transparency advantage of the proposed method over the conventional speech BWE methods [3, 5–7, 14, 39] was observed from the average MOS shown in Table 3. Moreover, the proposed method gives an MOS of 3.91 which being closer to the MOS for which the two signals sound identical shows that the quality of $X^1_{NB}(n)$ is almost identical to that of $X_{NB}(n)$. Obviously, the data embedding performed in the proposed speech BWE method has very little impact on perception.

5.1.3 Subjective Comparison of Bandwidth Extended NB Telephonic Speech with the Corresponding WB Speech Signal

A subjective test is carried out to evaluate the comparative performance of the bandwidth extended NB telephonic speech taken from NTIMIT database with the cor-

Table 2 Mean opinion scores (MOS)

Score	Instruction
1	Two signals are different
2	Two signals are similar, but the difference is easy to see
3	Two signals sound very similar, only little difference exists
4	Two signals sound identical

Table 3 Comparative performance in terms of average MOS

Method	MOS
Conventional data hiding [3]	3.02
Data hiding with Phonetic classification [5]	3.75
Conventional signal domain data hiding [6]	3.58
Conventional bit stream data hiding [7]	3.02
Conventional joint coding and data hiding [39]	3.53
Conventional WTSI [14]	3.88
Proposed method	3.91

Table 4 Subjective test results of the comparison between I and II

I	II
▷	20
◁	0
≈	0

responding WB speech signal taken from TIMIT database. The original WB speech is numbered as I; bandwidth extended NB telephonic speech (the reconstructed WB speech of proposed method for the telephone speech input) is denoted as II. The listeners were asked to do pairwise comparison of speech samples taken from I to II. They had to respond whether the first sample of the pair sounded better (▷), worse (◁) or equal (≈) when compared to second one. The results of comparing I–II are tabulated in Table 4. The number of listeners with a particular preference (▷ or ◁ or ≈) are provided in the Table in Arabic numbers. From Table 4, we observe that original WB speech is better than the bandwidth extended NB telephonic speech.

5.1.4 ITU-T Test Results

The speech samples used in the listening test were taken from the TIMIT database. Ten different sentences spoken by ten different male and female speakers of 2–2.5 s long were taken for evaluating the performance of conventional speech BWE methods [3, 5, 6] and the proposed method. Since the main application of the speech BWE technique is in mobile communications, listening test samples were prepared so that they simulated speech transmitted over a cellular telephone network. The test samples were high pass filtered with the MSIN filter, which approximates the input response of a mobile station and the sound level of each test sample was normalized to 26 dB below overloading [20]. These preprocessed test samples were then down sampled to the 8 kHz sampling rate and used as NB signal for conventional speech BWE methods [3, 5, 6] and the proposed method.

A subjective listening test is carried out to evaluate the comparative performance of conventional speech BWE methods [3, 5, 6] and the proposed method. The original WB speech taken from TIMIT database is numbered as I; the telephone speech taken from NTIMIT database is denoted as II; the composite NB speech is numbered as III, and the reconstructed WB speech is denoted as IV. The listeners were asked to do pairwise comparison of speech samples taken from I to IV. They had to respond whether the first sample of the pair sounded better (▷), worse (◁) or equal (≈) when compared to second one. The results of comparing I, II and III with the other signals are tabulated in Table 5a–c, respectively. The number of listeners with a particular preference (▷ or ◁ or ≈) are provided in the Table in Arabic numbers. From Table 5a, we observe that original WB speech is better than telephone speech and composite NB speech of conventional speech BWE methods [3, 5, 6] and the proposed method. It is also seen that, compared to the conventional speech BWE methods [3, 5, 6], the proposed scheme has a better WB reconstruction performance. Table 5b shows that, compared to the conventional speech BWE methods [3, 5, 6], the composite NB speech

Table 5 Subjective test results of the comparisons between (a) I and the others, (b) II and the others, and (c) III and IV

	I	II	III	IV
(a)				
Conventional data hiding [3]	▷	20	20	9
	◁	0	0	0
	≈	0	0	11
Data hiding with Phonetic classification [5]	▷	20	20	7
	◁	0	0	0
	≈	0	0	13
Conventional signal domain data hiding [6]	▷	20	20	6
	◁	0	0	0
	≈	0	0	14
Proposed method	▷	20	20	4
	◁	0	0	0
	≈	0	0	16
	II	III	IV	
(b)				
Conventional data hiding [3]	▷		5	3
	◁		9	10
	≈		6	7
Data hiding with Phonetic classification [5]	▷		4	2
	◁		11	12
	≈		5	6
Conventional signal domain data hiding [6]	▷		1	1
	◁		12	15
	≈		7	4
Proposed method	▷		0	0
	◁		14	19
	≈		6	1
		III	IV	
(c)				
Conventional data hiding [3]		▷		3
		◁		11
		≈		6
Data hiding with Phonetic classification [5]		▷		2
		◁		13
		≈		5
Conventional signal domain data hiding [6]		▷		1
		◁		16
		≈		3

Table 5 continued

	III	IV
Proposed method	▷	0
	◁	18
	≈	2

of proposed method is better than telephone speech and also that there is a clear quality improvement of the reconstructed WB speech of proposed method over the telephone speech. From Table 5c, we observe that, compared to the conventional speech BWE methods [3, 5, 6], the reconstructed WB speech of the proposed method is better than composite NB speech.

5.2 Objective Quality Evaluation

To further evaluate the proposed method, the same data base used in the subjective listening tests is evaluated using the objective measures. The quality of reconstructed UB speech was assessed using LSD measure [3], and perceptual transparency was assessed using the ITU-T PESQ tool [21]. The robustness of embedded information against quantization and channel noises was assessed using BER. The quality of reconstructed WB speech was assessed using the WB-PESQ measure [22] recommended by ITU-T. The spectrogram analysis was employed to compare the performance of proposed and conventional speech BWE methods for fricatives. The comparison of pitch contours before and after proposed method was employed. The comparison of bandwidth extended NB telephonic speech with the corresponding WB speech signal and the comparison of telephone speech with the composite NB speech was also employed.

5.2.1 Comparison of Original and Reconstructed UB Speech

LSD measure, which is based on comparing the short-time spectral envelopes, is used in this paper to assess the perceptual similarity of the true and the reconstructed UB signals. The LSD measure is computed using the formula

$$\text{LSD} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(20 \log_{10} \frac{G}{|A(e^{jw})|} - 20 \log_{10} \frac{\hat{G}}{|\hat{A}(e^{jw})|} \right)^2 dw \quad (47)$$

where G and $\frac{1}{A(e^{jw})}$ are the gain and the true UB spectral envelope, respectively; and \hat{G} and $\frac{1}{\hat{A}(e^{jw})}$ are the gain and the reconstructed UB spectral envelope, respectively. The spectral envelopes are calculated based on linear prediction for short frames of 20 ms long. Finally, the mean of the LSD measure is calculated over all frames of the signal to be evaluated. A smaller value of LSD indicating a superior quality of

Table 6 Comparative performance in terms of average LSD

Method	LSD
Conventional data hiding [3]	12.17
Data hiding with Phonetic classification [5]	10.46
Conventional signal domain data hiding [6]	6.3
Conventional bit stream data hiding [7]	5.89
Conventional joint coding and data hiding [39]	5.05
Conventional WTSI [14]	2.9
Proposed method	2.43

Table 7 Comparative performance in terms of average NB-PESQ

Method	NB-PESQ
Conventional data hiding [3]	3.21
Data hiding with Phonetic classification [5]	3.96
Conventional signal domain data hiding [6]	3.75
Conventional bit stream data hiding [7]	3.17
Conventional joint coding and data hiding [39]	3.68
Conventional WTSI [14]	3.98
Proposed method	4.05

reconstructed UB signal. Table 6 presents the resultant average LSD over all samples of conventional speech BWE methods [3, 5–7, 14, 39] and proposed method with a μ -law channel model.

From Table 6, we observe that the proposed method consistently outperforms the conventional speech BWE methods [3, 5–7, 14, 39]. This happens because the number of error bits is decreased due to the suppression of channel noise, quantization noise and channel spectral distortion, and will give a smaller error in embedded information extracting. Moreover, the proposed method gives a low LSD of 2.43 which shows that the quality of the reconstructed WB speech $X^1_{WB}(n)$ of the proposed method almost reaches the quality level of original WB speech $X_{WB}(n)$. These LSD values support the good WB performance of the proposed method which was already found in the subjective tests. The proposed method gives an LSD of 2.57 with AWGN channel model.

5.2.2 Perceptual Transparency

The NB-PESQ measure is used to evaluate the perceptual transparency by providing $X_{NB}(n)$ (reference speech) and $X^1_{NB}(n)$ (degraded version of speech) as inputs. Here, the speech quality is rated using NB-PESQ measure by comparing $X_{NB}(n)$ with $X^1_{NB}(n)$. The PESQ returns a score from 0.5 to 4.5, with higher scores indicating superior quality. Table 7 presents the resultant average NB-PESQ score over all samples of conventional speech BWE methods [3, 5–7, 14, 39] and proposed method. A clear

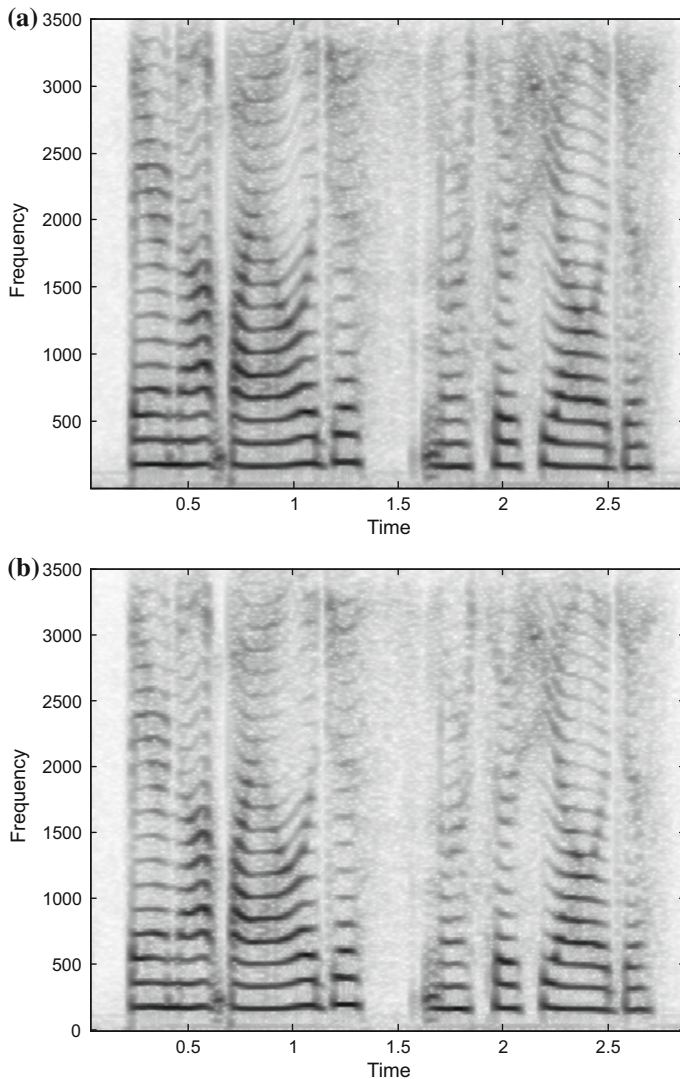


Fig. 4 **a** NB speech signal. **b** Composite NB speech signal

perceptual transparency improvement of the proposed method over the conventional speech BWE methods [3,5–7,14,39] was observed from the average NB-PESQ scores shown in Table 7. The proposed method gives a PESQ score of 4.05 which shows that excellent perceptual transparency of the proposed method which was already found in the subjective listening tests.

The spectrograms of the NB signal $X_{NB}(n)$ and the composite NB speech signal $X_{NB}^1(n)$ are shown, respectively, in Fig. 4a, b. It is clear from the figures that $X_{NB}(n)$ and $X_{NB}^1(n)$ are almost indistinguishable.

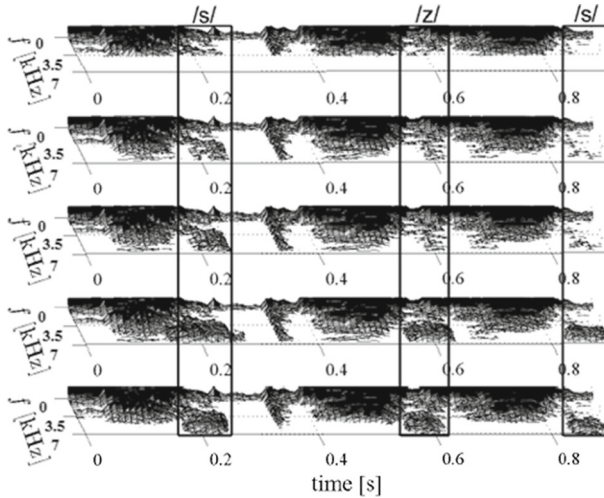


Fig. 5 Spectrograms from *top to bottom*: **a** NB speech, **b** conventional data hiding, **c** data hiding with phonetic classification, **d** proposed method, **e** original WB speech

5.2.3 Robustness of Embedded Information

Next, we consider the effect of noise corruption. AWGN is added to composite NB signal $X_{NB}^1(n)$, with a SNR of 35 dB. The BER is used as a performance measure. We used 6 bits to encode the entry index of the VQ codebook; i.e., the VQ codebook has a size of $2^6 = 64$. The SS length is fixed at 8. A smaller value of BER indicates a superior quality of reconstructed signal. It is observed that with SNR of 35 dB, the obtained BER is 4.52×10^{-4} , which shows that the embedded information was successfully recovered by employing DS-CDMA technique.

Although the μ -law coding causes distortions to the embedded information, the obtained BER after applying μ -law coding to $X_{NB}^1(n)$ is 1.31×10^{-4} , which shows that the embedded information was properly retrieved by employing DS-CDMA technique.

5.2.4 Spectrogram Analysis

The performance of the proposed, conventional data hiding [3] and data hiding with phonetic classification [5] techniques for fricatives were compared in terms of spectrogram analysis. The spectrograms of the utterance “less poisonous” reconstructed by different methods were illustrated in Fig. 5. The fricatives are marked on the top of Fig. 5. The upper plot 5a depicts the spectrogram of NB speech, whereas the lower plot 5e depicts the spectrogram of original WB speech. The middle plots 5b–d show the spectrograms of the conventional data hiding [3], data hiding with phonetic classification [5] and proposed methods. Note the improvement of 5c over 5b at the first instance of /s/. A significant improvement of proposed method can be reported for graph 5d, which is—for all fricative instances /s/ and /z/—very close to graph 5e.

Table 8 Comparative performance in terms of average WB-PESQ

Method	WB-PESQ
Conventional data hiding [3]	2.62
Data hiding with Phonetic classification [5]	2.85
Conventional signal domain data hiding [6]	3.79
Conventional bit stream data hiding [7]	3.83
Conventional joint coding and data hiding [39]	3.19
Conventional WTSI [14]	4.05
Proposed method	4.16

5.2.5 WB Speech Quality

The speech samples used in the WB-PESQ measure [22] were taken from the TIMIT database. Ten different sentences spoken by fifteen different male and female speakers of 2–2.5 s long were taken for evaluating the performance of conventional speech BWE methods [3,5–7,14,39] and the proposed method. The WB-PESQ measure is used to evaluate the quality of reconstructed WB speech by providing original WB speech taken from TIMIT database $X_{WB}(n)$ and reconstructed WB speech $X^1_{WB}(n)$ as inputs. Here, the speech quality is rated using WB-PESQ measure by comparing $X_{WB}(n)$ with $X^1_{WB}(n)$. Table 8 presents the resultant average WB-PESQ over all samples of conventional speech BWE methods [3,5–7,14,39] and proposed method. A clear quality improvement of the proposed method over the conventional speech BWE methods [3,5–7,14,39] was observed from the average WB-PESQ scores shown in Table 8. The proposed method gives a PESQ score of 4.16 which shows that excellent reconstructed WB speech quality of the proposed method which was already found in the subjective listening tests.

5.2.6 Comparison of Pitch Contours

The time domain waveforms, pitch contours and spectrograms of original WB speech $X_{WB}(n)$ and reconstructed WB speech $X^1_{WB}(n)$ are shown, respectively, in Figs. 6 and 7. It is clear from the figures that the pitch contours of $X_{WB}(n)$ and $X^1_{WB}(n)$ are almost same. This happens because the pitch (F0) information is present in the low-frequency region (below 300 Hz) of a NB signal and the spreaded spectral envelope parameters of the extendedband signal are embedded in the low-amplitude high-frequency part of the magnitude spectrum of NB speech without altering the low-frequency region.

5.2.7 Objective Comparison of Telephone Speech and Composite NB Speech

Ten different sentences spoken by fifteen different male and female speakers of 2–2.5 s long were taken for evaluating the performance of conventional speech BWE methods [3,5–7,14,39] and the proposed method. The NB-PESQ measure is used to evaluate the comparative performance between telephone speech taken from the NTIMIT database

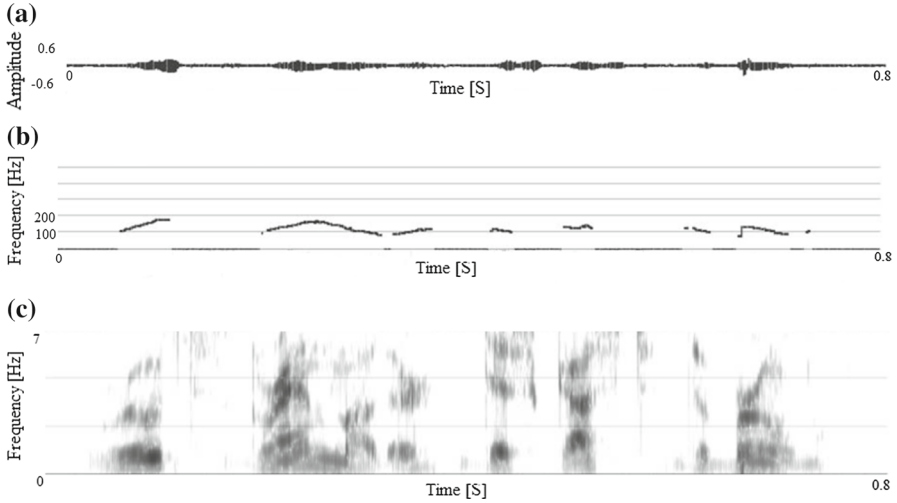


Fig. 6 **a** Time domain waveform of original WB speech. **b** Pitch contour of original WB speech. **c** Spectrogram of original WB speech

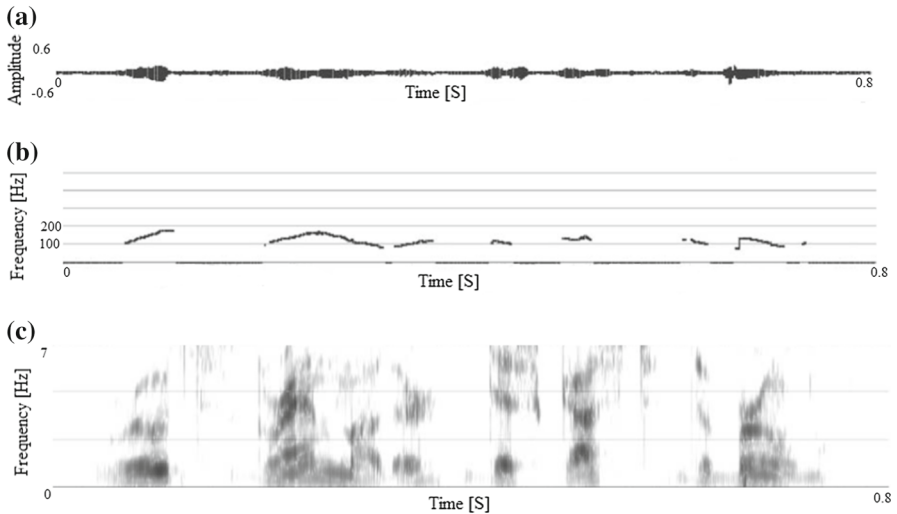


Fig. 7 **a** Time domain waveform of reconstructed WB speech. **b** Pitch contour of reconstructed WB speech. **c** Spectrogram of reconstructed WB speech

and composite NB speech by providing telephone speech and composite NB speech as inputs. Here, the speech quality is rated using NB-PESQ measure by comparing telephone speech with composite NB speech. Table 9 presents the resultant average NB-PESQ score over all samples of conventional speech BWE methods [3, 5–7, 14, 39] and proposed method. Compared to the conventional speech BWE methods [3, 5–7, 14, 39], the composite NB speech of proposed method is better than telephone speech which was already found in the subjective listening tests.

Table 9 Comparative performance in terms of average NB-PESQ

Method	PESQ
Conventional data hiding [3]	1.67
Data hiding with Phonetic classification [5]	1.95
Conventional signal domain data hiding [6]	2.83
Conventional bit stream data hiding [7]	2.48
Conventional joint coding and data hiding [39]	2.79
Conventional WTSI [14]	2.96
Proposed method	3.18

5.2.8 Objective Comparison of Bandwidth Extended NB Telephonic Speech with the Corresponding WB Speech Signal

The WB-PESQ measure is used to evaluate the quality of the bandwidth extended NB telephonic speech (the reconstructed WB speech of proposed method for telephone speech input). Here, the speech quality is rated using WB-PESQ measure by comparing original WB speech $X_{WB}(n)$ with the bandwidth extended NB telephonic speech $X^1_{WB}(n)$. The proposed method gives a PESQ score of 2.67 which shows that poor bandwidth extended NB telephonic speech quality of the proposed method.

6 Conclusion

A novel NB speech BWE technique using magnitude spectrum data hiding for extending the bandwidth of the existing NB telephone networks is proposed. The spreaded spectral envelope parameters of extendedband signal are embedded in the low-amplitude high-frequency part of the magnitude spectrum of NB signal at the transmitter. The embedded information is extracted at the receiving end to reconstruct the wideband speech signal.

DS-CDMA technique is employed to increase the robustness of the embedded extendedband signal to quantization and channel noises by spreading the spectral envelope parameters by multiplying them with SSs and then adding them up together to provide the embedded information. The embedded information can be reliably recovered by using a multiuser detector. The experimental results show that the proposed method is found to be robust to quantization and channel noises. CCR and LSD test results indicate that there is a clear speech quality improvement of the proposed method over the conventional data hiding, data hiding with phonetic classification, conventional signal domain data hiding, conventional joint coding and data hiding, conventional bit stream data hiding and conventional WTSI techniques. The MOS test value obtained for the proposed method indicate that the method embeds the UB information more transparently compared to the conventional data hiding, data hiding with phonetic classification, conventional signal domain data hiding, conventional joint coding and data hiding, conventional bit stream data hiding and conventional WTSI techniques. The proposed method is demonstrated to produce a much better

quality speech signal than the conventional data hiding, data hiding with phonetic classification, conventional signal domain data hiding, conventional joint coding and data hiding, conventional bit stream data hiding and conventional WTSI techniques. Hence, it is suitable for extending the bandwidth of the existing telephone networks without making changes to the telephone networks.

References

1. S. Andreas, P. Ted, A. Venkatraman, *Audio Signal Processing and Coding* (Wiley-Interscience Publication, USA, 2006)
2. P. Bauer, T. Fingscheidt, An HMM based artificial bandwidth extension evaluated by cross-language training and test, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, April 2008, pp. 4589–4592
3. S. Chen, H. Leung, Artificial bandwidth extension of telephony speech by data hiding, in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS 2005)*, Kobe, Japan, May 2005, pp. 3151–3154
4. S. Chen, H. Leung, Concurrent data transmission through analog speech channel using data hiding. *IEEE Signal Process. Lett.* **12**(8), 581–584 (2005)
5. 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 (ICASSP)*, Honolulu, HI, April 2007, vol. 4 (2007), pp. 593–596
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.* **2013**(10), 1–8 (2013)
8. E.H. Dinan, E.H. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks. *IEEE Commun. Mag.* **36**(9), 48–54 (1998)
9. H. Ding, Wideband audio over narrowband low-resolution media, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, March 2004, pp. 489–492
10. J. Epps, W.H. Holmes, A new technique for wideband enhancement of coded narrowband speech, in *Proceedings of IEEE Workshop on Speech Coding*, Porvoo, June 1999, pp. 174–176
11. European Telecommunications Standards Institute (ETSI) Standard, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 V1.1.2, April 2000
12. W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd edn. (Wiley, New York, 1970)
13. J.S. Garofolo, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database* (National Institute of Standards and Technology (NIST), Gaithersburg, 1988)
14. B. Geiser, P. Jax, P. Vary, Artificial bandwidth extension of speech supported by watermark-transmitted side information, in *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, September 2005, pp. 1497–1500
15. B. Geiser, P. Vary, Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, April 2007, vol 4 (2007), pp. 533–536
16. A. Goldsmith, *Wireless Communications* (Cambridge University Press, New York, 2005)
17. E. Hansler, G. Schmidt, *Speech and Audio Processing in Adverse Environments* (Springer, Berlin, 2008)
18. L. Hanzo, F.C.A. Somerville, J.P. Woodard, *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels* (IEEE Press, Hoboken, 2001)
19. International Telecommunications Union, Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, August 1996
20. International Telecommunications Union, Software tools for speech and audio coding standardization, ITU-T Rec. G.191, September 2005

21. 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, ITU-T Recommendation P.862, February 2001
22. International Telecommunications Union, Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs, ITU-T Recommendation P.862.2, November 2005
23. B. Iser, W. Minker, G. Schmidt, *Bandwidth Extension of Speech Signals* (Springer, New York, 2008)
24. P. Jax, Enhancement of bandlimited speech signals: algorithms and theoretical bounds. Ph.D. thesis, RWTH Aachen University, 2002
25. 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 (ICASSP)*, Orlando, FL, USA, May 2002, vol 1 (2002), pp. 237–240
26. P. Jax, P. Vary, On artificial bandwidth extension of telephone speech. *Signal Process.* **83**(8), 1707–1719 (2003)
27. 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)
28. Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design. *IEEE Trans. Commun.* **28**(1), 84–95 (1980)
29. Y. Nakatoh, M. Tsushima, T. Norimatsu, Generation of broadband speech from narrowband speech using piecewise linear mapping, in *Proceedings of EUROSPEECH*, Rhodes, Greece, September, 1997, pp. 1643–1646
30. M. Nilsson, W.B. Kleijn, Avoiding overestimation in bandwidth extension of telephony speech, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001, vol 2 (2001), pp. 869–872
31. J.G. Proakis, *Digital Communications*, 2nd edn. (McGraw-Hill, New York, 1989)
32. H. Pulakka, P. Alku, Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Melspectrum. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2170–2183 (2011)
33. H. Pulakka, U. Remes, K. Palomaki, M. Kurimo, P. Alku, Speech bandwidth extension using gaussian mixture model-based estimation of the highband Mel spectrum, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, May 2011, pp. 5100–5103
34. Y. Qian, P. Kabal, Dual-mode wideband speech recovery from narrowband speech, in *Proceedings of EUROSPEECH 2003*, Geneva, September 2003, pp. 1433–1436
35. T. Rabie, D. Guerchi, Magnitude spectrum speech hiding, in *Proceedings of IEEE International Conference on Signal Processing and Communications (ICSPC 2007)*, Dubai, November 2007, pp. 1147–1150
36. R. Hu, V. Krishnan, D.V. Anderson, Speech bandwidth extension by improved codebook mapping towards increased phonetic classification, in *Proc. INTERSPEECH 2005*, Lisbon, Portugal, September 2005, pp. 1501–1504
37. A.H. Sayed, *Adaptive Filters* (Wiley, Hoboken, 2008)
38. W. Strange, T.R. Edman, J.J. Jenkins, Acoustic and phonological factors in vowel identification. *J. Exp. Psychol. Hum. Percept. Perform.* **5**(4), 643–656 (1979)
39. P. Vary, B. Geiser, Steganographic wideband telephony using narrowband speech codecs, in *Proceedings of Asilomar Conference on Signals, Systems, and Computers (ACSSC 2007)*, Pacific Grove, CA, November 2007, pp. 1475–1479
40. S. Vaseghi, E. Zavarzheh, Q. Yan, Speech bandwidth extension: Extrapolations of spectral envelop and harmonicity quality of excitation, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, May 2006, pp. 844–847