International Journal of Speech Technology (2022) 25:697–706
https://doi.org/10.1007/s10772-022-09980-x

Discrete cosine transform-based data hiding for speech bandwidth extension

Sunil Kumar Koduri[1] · Kishore Kumar T[1]

Received: 11 March 2021 / Accepted: 28 May 2022 / Published online: 24 June 2022

Abstract

The limited narrow frequency range of 300–3400 Hz used in public switched telephone networks causes a significant reduction of speech quality. To address this drawback, a new robust transform-domain speech bandwidth extension method is proposed in this paper. The method uses the discrete Cosine transform-based data hiding (DCTBDH) technique to provide a better-quality wideband speech signal. The spectral envelope parameters are extracted from the high-frequency components of speech signal existing above narrowband, which are then spread by using spreading sequences, and are embedded within the DCT coefficients of narrowband signal. A better-quality wideband signal is reconstructed using the extracted embedded information at the receiver end. In simulations, the high-quality wideband speech was obtained from speech transmitted over a public switched telephone network. The spectral envelope parameters of the high-frequency components of the speech signal are transparently embedded with a mean square error of 5.78×10^{-4}. In a mean opinion score (MOS) listening test, we verified that the proposed method yields improved perceptual transparency compared to conventional methods of about 0.21 points on the MOS scale. The log spectral distortion value obtained was 2.2248 which showed that the proposed technique yields an improved quality of speech signal compared to conventional methods.

Keywords Discrete Cosine transform-based data hiding · Linear prediction · Speech bandwidth extension · Speech quality · Spread spectrum

1 Introduction

Most of the traditional telephone networks allow only a narrowband (NB) signal which is band-limited to 300–3400 Hz. Usually, human speech contains frequencies far beyond the NB frequency range. Thus, the transmission of human speech through the networks leads to the muffled sound and poor-quality telephony speech. Wideband (WB) speech transmission in the range of 50–7000 Hz would be desirable for better speech quality. To allow WB speech services, the essential changes required within the network infrastructure are quite expensive and time-taking (Jax & Vary, 2006). This is happening to be a major hurdle for the transmission of high-quality speech in telephone networks. Therefore, it is very important to enable WB speech transmission using speech bandwidth extension (SBE) techniques to enhance the quality of speech (Jax, 2002).

Artificial bandwidth extension (ABE) is one among various methods of SBE which improve the quality and intelligibility of telephony speech (Jax, 2002). In this approach, the out-of-band information i.e. the frequencies below 300 Hz and above 3400 Hz are estimated from the NB signal. Excitation signal and spectral envelop (SPENV) are estimated by most ABE techniques which are used to regenerate the out-of-band signal. Different approaches for extension of excitation signal are presented in (Jax, 2002; Prasad & Kishore Kumar, 2016). Different techniques for estimating WB SPENV are presented in Prasad & Kishore Kumar (2016), Zhen-Hua et al. (2018), Bong-Ki et al. (2018), Abel and Fingscheidt (2017), Yingwue et al. (2016). In Yuanjie Dong et al. (2020), a time–frequency network with channel attention and non-local modules is used for SBE. Latent representation learning for ABE using a conditional variational auto-encoder is presented to enhance the speech quality in Pramod Bachhav et al. (2019). The time-domain multi-scale

✉ Sunil Kumar Koduri
sunil.veena10@gmail.com

Kishore Kumar T
kishorefr@gmail.com

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

fusion neural network approach for improving the performance of SBE is presented in Xiang Hao et al. (2020). SBE using a conditional generative adversarial network with discriminative training is introduced in Jonas Sautter et al. (2019) to reconstruct the better-quality speech signal. The audio signal of better quality is regenerated in Mathieu and Felix (2020) using audio bandwidth extension aided by dilated convolutional neural network approach. In Kyoungjin and Joon et al. (2020), a deep neural network ensemble approach for reducing artificial noise in SBE is introduced. A waveform-based method for SBE that uses a deep three-way split summation FFTNet architecture is proposed in Berthy Feng et al. (2019). In Johannes and Tim (2019), a time-domain ABE technique towards low-frequency band by a sinusoidal synthesis of missing harmonics is presented to enhance the quality of the reconstructed WB signal. A WaveNet-based model conditioned on a log-mel spectrogram representation of NB speech to reconstruct the better quality speech signal is proposed in Archit Gupta et al. (2019). However, traditional ABE methods are suffering from reconstructing WB speech with high quality under all conditions (Jax & Vary, 2002).

Compared to ABE techniques, a WB speech with high quality is reconstructed when the out-of-band information is transmitted by hiding it in the NB signal using data hiding methods (Jax & Vary, 2006). Several techniques for SBE using data hiding are proposed in the state-of-the-art literature. An SBE technique is proposed in Jax and Vary (2006) to embed the encoded SPENV parameters of the lost speech frequency components within the NB signal. A better-quality WB signal is reconstructed at the receiver end using the embedded information. A much better-quality WB signal over (Chen & Leung, 2005) has been reconstructed in Chen & Leung (2007), where the spectral envelope parameters (SPEVPS) are efficiently encoded using phonetic classification. The pitch-scaled frequencies of the out-of-band signal are embedded into the unused frequencies of traditional telephony speech to enhance the quality of reconstructed WB speech in Geiser and Vary (2013). The WB signal of better quality is regenerated in Bhatt and Kosta (2015), Geiser and Vary (2007), Kosta (2016) using joint source coding and data hiding technique. High-quality WB signal is reconstructed in Prasad and Kishore Kumar (2017), Sunil Kumar and Kishore Kumar (2019) using various frequency-domain data hiding techniques. The enhancement in the quality of reconstructed WB speech is achieved by restoring the hidden audible components of the out-of-band signal (Chen & Leung, 2007). The SPEVPS of an out-of-band signal is embedded into the NB signal bitstream to improve the quality of reconstructed WB speech in Chen et al. (2013). The WB signal of better quality is regenerated in Sagi and Malah (2007) using the quantization-based watermarking technique.

SBE techniques with data hiding are expected to deliver high-quality composite narrowband (CNB) alongside reconstructed wideband (RWB) signals. Also, these techniques must be able to handle issues pertaining to quantization and channel noises. Nevertheless, most of the traditional techniques fail to provide high-quality CNB and RWB signals (Bhatt & Kosta, 2015; Chen & Leung, 2007; Chen & Leung, 2007; Chen et al., 2013; Geiser & Vary, 2007, 2013; Jax & Vary, 2006; Kosta, 2016; Prasad & Kishore Kumar, 2017; Sagi & Malah, 2007; Sunil Kumar & Kishore Kumar, 2019). Also, they are less robust to channel and quantization noises. Thus, developing a novel SBE technique using data hiding is essential to improve the quality of CNB and RWB signals and more robust to channel and quantization noises.

An audio steganography technique is presented in Kanhe and Aghila (2016), used discrete Cosine transform (DCT)-based data hiding technique to insert the secret message signal in DCT coefficients of a host speech signal without degrading the perceptual quality of the host signal. It was shown that this approach is producing a stego signal that is indistinguishable from the host signal while being able to reliably recover the secret message signal at the receiver end without any degradation in quality.

A novel SBE algorithm using the discrete Cosine transform-based data hiding technique (Kanhe & Aghila, 2016) is proposed to embed the parameters of the lost speech frequency components within the DCT coefficients of the NB signal. These hidden parameters are retrieved at the receiver side to produce a better-quality WB signal by combining the missing speech signal that was transmitted through the DCT coefficients and the NB signal. The proposed scheme uses the real missing speech information instead of its estimation which makes the reconstruction of the WB speech more accurate compared to the conventional ABE methods. Furthermore, the proposed method is compatible with conventional NB terminal equipment, e.g., a plain ordinary telephone set. In other words, conventional NB receivers can still access the NB speech properly without additional hardware, while a customized receiver can extract the embedded information and provide WB signal with much better quality.

The telephone network channel introduces channel and quantization noises. Techniques proposed in Chen and Leung (2007), Geiser et al. (2005), Jax and Vary (2006) for SBE are considered only the quantization noise ignoring the channel noise. The quantization noise and channel noise effects are considered in this paper. The spread spectrum technique (Hasan et al. 1998) is used in this work for retrieving the embedded information as it is claimed to be more robust against quantization and channel noises. In particular, each parameter to be inserted is spread by multiplying with a particular spreading sequence. The embedded information is then formed by adding the spread signals. Due to orthogonality among spreading sequences, the embedded information is retrieved reliably by using a correlator.

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 (Dianan & Jabbari, 1998; Goldsmith, 2006). 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.

The paper is ordered as follows. In Sect. 2, the DCTBDH method for SBE is introduced. The subjective and objective analyses are discussed in Sect. 3. Finally, in Sect. 4, conclusions are summarized.

2 DCT-based data hiding technique for SBE

2.1 Transmitter

The proposed transmitter is shown in Fig. 1. A speech signal designated as WB $S_{wb}(n)$ is sampled at a frequency of 16 kHz. This signal is further fragmented to form a low band signal using a low pass filter (LPF) and a high band signal using a high pass filter (HPF) respectively. The LPF extracts speech signal information that is present between 0 and 4 kHz and is designated as low band signal while HPF extracts speech information that is present between 4 and 8 kHz designated as high band signal. The LPF output is decimated by a factor of two to produce an NB signal $S_{nb}(n)$. The high band signal is decimated to produce an upper band (UB) signal $S_{eb}(n)$. Therefore, 8 kHz is the sampling frequency of $S_{nb}(n)$ and $S_{eb}(n)$.

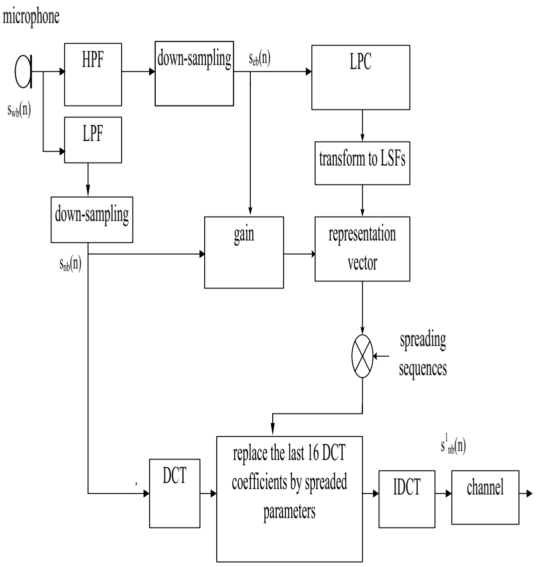

Fig. 1 Proposed transmitter

To imperceptibly embed $S_{eb}(n)$ in to $S_{nb}(n)$, the number of parameters that represents $S_{eb}(n)$ is minimized. Here, the linear predictive analysis (Hanzo et al., 2001) is employed to accomplish this target. Linear predictive analysis is based on the source-filter model of speech production. The linear predictive coefficients (LPC) are the reciprocal of the autoregressive filter coefficients. The LPC which correspond to the spectral envelope of $S_{eb}(n)$ are denoted as$b_i(i = 1, \ldots, 10)$, where i is the order of filter, which are found by using the Levinson-Durbin algorithm. The small variation in LPC results in substantial distortions when reconstructing $S_{eb}(n)$; hence LPC are modified into line spectral frequencies (LSF) (Hanzo 2001). Also, the gain of$s_{eb}(n)$, denoted withG_r, has to be embedded since synthesized UB speech has to be scaled to an appropriate energy to evade over-estimation (Nilsson & Kleijn, 2001). Thus, the representation vector which represents $S_{eb}(n)$ is formed by combining LSF and gain, i.e.,$R = \left[LSF_1, LSF_{2,\ldots\ldots,} LSF_{10}, G_r\right]$.

The excitation parameters of $S_{eb}(n)$ are not embedded to reduce the number of parameters to be hidden. This is because, the ear is not very sensitive to distortions of the excitation signal at above NB (Jax & Vary, 2003). Thus, estimating the excitation of $S_{eb}(n)$ at the receiver from $S_{nb}(n)$ is well-suited for the reconstruction performance.

All the parameters of R is denoted with D_i. One among all the parameters of R is then denoted with D_{i0}. Every parameter of R to be embedded is spread by multiplying it with a particular spreading sequence, i.e.,$D_i \cdot p^{\rightarrow i}, 1 \le i \le Q$. The hidden data is then produced by adding all of these spreading vectors and is given by

$$V(g) = \sum_{i=1}^{Q} D_i p^i(g) \tag{1}$$

where g^{th} element of $p^{\rightarrow i}$ represented by $p^i(g)$. DCT is then applied on NB signal $S_{nb}(n)$ and can be expressed as

$$S_{nb}(k) = w(k) \sum_{n=0}^{N-1} S_{nb}(n) \cos \frac{(2n+1)k\pi}{2N}, k = 0toN - 1 \tag{2}$$

where

$$w(k) = \sqrt{\frac{1}{N}}\ if\ k = 0, w(k) = \sqrt{\frac{2}{N}}\ otherwise$$

The last 16 coefficients of the DCT coefficients are replaced by $V(g)$ resulting in a CNB signal spectrum (Kanhe & Aghila, 2016). To transform back the CNB signal spectrum to time-domain representation, inverse discrete cosine transform (IDCT) is applied on the CNB signal spectrum and can be expressed as

$$S_{nb}(n) = \sum_{k=0}^{N-1} w(k)S_{nb}(k)\cos\frac{(2n+1)k\pi}{2N}, n = 0 \quad to \quad N-1 \tag{3}$$

Thus, a CNB signal $S^1_{nb}(n)$ is produced so that it can be communicated to the receiver on a telephone network channel.

A synchronization sequence like 111….0.11 is added after every frame of the CNB signal to achieve frame synchronization (ETSI ES 201 2000) between the transmitter and receiver. The arrival of a new frame of the CNB signal at the receiver is indicated by the reception of a synchronization sequence.

2.2 Receiver

The proposed receiver is shown in Fig. 2. The CNB signal received through a telephone network channel is noisy. Assume that the received signal is represented by $\hat{\mathrm{S}}^1_{nb}(n)$ i.e., $\hat{\mathrm{S}}^1_{nb}(n) = S^1_{nb}(n) + e$. Where e represents the combination of channel and quantization noises. The conventional phone terminal treats $\hat{\mathrm{S}}^1_{nb}(n)$ as an ordinary signal. The NB signal quality is not noticeably degraded since there is a very small perceived difference between $S^1_{nb}(n)$ and $\hat{\mathrm{S}}^1_{nb}(n)$. Retrieval of the embedded data requires applying DCT on the CNB signal to obtain the DCT coefficients.

The spread parameters are then obtained from the last 16 DCT coefficients (Kanhe & Aghila, 2016) and a correlator is used to de-spread these parameters. Assuming a particular $\underset{\wedge}{\mathrm{D}}_i$ is denoted as $\underset{\wedge}{\mathrm{D}}_{io}$ to be retrieved, the correlation can be expressed as

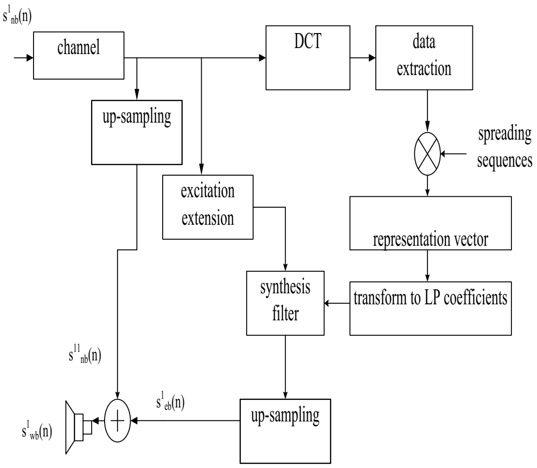

Fig. 2 Proposed receiver

$$\underset{\wedge}{\mathrm{D}}_{io} = \frac{1}{Q}\sum_{g=1}^{Q} \underset{\cdot}{\mathrm{V}}(g)\, p^{io}(g) \tag{4}$$

where $\underset{\cdot}{\mathrm{V}}(g)$ represents noisy $V(g)$ and is given by

$$\underset{\cdot}{\mathrm{V}}(g) = V(g) + \bar{\mathrm{e}}(g) \tag{5}$$

Equation (5) is substituted into Eq. (4), so that we have

$$\underset{\wedge}{\mathrm{D}}_{io} = \frac{1}{Q}\sum_{g=1}^{Q} \underset{\cdot}{\mathrm{V}}(g)\, p^{io}(g)$$

$$= \frac{1}{Q}\sum_{g=1}^{Q} p^{io}(g)\left(\sum_{i=1}^{Q} \hat{D}_i p^i(g) + \overline{e}(g)\right)$$

$$= \frac{1}{Q}\sum_{g=1}^{Q} p^{io}(g) \times \left(\hat{D}_{io}p^{io}(g) + \sum_{i\neq io} \hat{D}_i p^i(g) + \overline{e}(g)\right)$$

$$= \hat{D}_{io} + \frac{1}{Q}\sum_{g=1}^{Q}\sum_{i\neq io} \hat{D}_i p^i(g)p^{io}(g) + \frac{1}{Q}\sum_{g=1}^{Q} p^{io}(g)\overline{e}(g) \tag{6}$$

The PN sequences are orthogonal. i.e.

$$\sum_{g=1}^{Q} p^i(g)p^{io}(g) = 0$$

where $i \neq io$. Therefore

$$\sum_{g=1}^{Q}\sum_{i\neq io} \hat{D}_{io}p^i(g)p^{io} = \sum_{i\neq io} \hat{D}_{io} \sum_{g=1}^{Q} p^i(g)p^{io}(g) = 0 \tag{7}$$

Also, since there was no correlation between $p^{io}(g)$ and $\overline{e}(g)$ i.e.

$$\frac{1}{Q}\sum_{g=1}^{Q} p^{io}(g)\bar{\mathrm{e}}(g) = 0 \tag{8}$$

when $Q \to \infty$. Equations (7) and (8) are substituted into Eq. (6), thus we have

$$\underset{\wedge}{\mathrm{D}}_{io} = \check{\mathrm{D}}_{io} \tag{9}$$

This reveals that the parameters which represent $\hat{\mathrm{S}}_{eb}(n)$ can be effectively recovered from using the SS technique and then the LPC are obtained from LSF. Meanwhile, NB residual signal is obtained by inverse filtering $\hat{\mathrm{S}}^1_{nb}(n)$ using LPC of $\hat{\mathrm{S}}^1_{nb}(n)$ and then obtain the UB excitation signal by extending the NB residual signal. The UB signal $\hat{\mathrm{S}}_{eb}(n)$ that was embedded is synthesized by exciting the synthesis filter described by the recovered LPC by a UB excitation signal. The received CNB and reconstructed UB signals are sampled at an 8 kHz sampling rate. These signals are then interpolated by a factor of two. $S^1_{eb}(n)$, represents interpolated $\hat{\mathrm{S}}_{eb}(n)$ signal. The interpolated CNB ($S^{11}_{nb}(n)$) and UB $S^1_{eb}(n)$ signals are added up for reproducing a WB signal ($S^1_{wb}(n)$) of good quality.

3 Experimental results

The speech utterances used for the performance evaluations of traditional and proposed SBE techniques were obtained from the TIMIT database (Garofalo et al., 2013). The evaluations were done by taking thirty different speech utterances which were spoken by thirty female and male speakers. The performance assessment of the methods was done by considering the subjective as well as objective measures. Each speech signal was split to form frames of 20 ms long and between frames an overlap of 10 ms was maintained. Each frame was processed individually. Existing SBE algorithms like data hiding (Chen & Leung, 2005), phonetic classification (Chen & Leung, 2007), audio watermark (Chen et al., 2013), steganographic WB telephony (Bhatt & Kosta, 2015), magnitude spectrum data hiding (Prasad & Kishore Kumar, 2017), and transform domain data hiding (Sunil Kumar & Kishore Kumar, 2019) were compared with the proposed method. Additive white Gaussian noise (AWGN) and μ-law channel models were used for analysis.

3.1 Subjective listening test results

The perceptual transparency was assessed with the mean opinion score (MOS) test (Chen & Leung, 2005, 2007). The subjective comparison between WB, CNB, NB, and RWB signals was also employed (Prasad & Kishore Kumar, 2017). Each person is made to hear the speech utterances through headphones in a silent chamber. An evaluation was done using a predefined scale by examining participant's views on speech sounds. Thirty persons have participated in the tests.

3.1.1 Perceptual transparency

The CNB and NB signals have to be similar sounds. The perceptual transparency was assessed with the MOS test. While comparing CNB and NB signals the listener comes out with a decision in terms of MOS as given in Table 1. The average MOS values of traditional (Bhatt & Kosta, 2015; Chen & Leung, 2005, 2007; Chen et al., 2013; Prasad & Kishore Kumar, 2017; Sunil Kumar & Kishore Kumar, 2019) and the proposed techniques are given in Table 2. The proposed technique gave a MOS value of 3.99 which indicates that the proposed technique has excellent perceptual transparency over the traditional techniques. The proposed technique gave a MOS value of 3.99 which was almost near the standard MOS value of 4 which indicates that CNB and NB signals were more or less identical.

Table 1 MOS

Score	Instruction
1	NB and CNB signals sound different
2	Observable difference between NB and CNB Signals
3	Minute difference between NB and CNB Signals
4	NB and CNB Signals sound alike

Table 2 Result of MOS

Method	Value
Data hiding (Chen & Leung, 2005)	2.89
Phonetic classification (Chen & Leung, 2007)	3.07
Audio watermark (chen et al., 2013)	3.18
Side information (Bhatt & Kosta, 2015)	3.54
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	3.63
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	3.74
Proposed method	3.99

3.1.2 Subjective comparisons between WB, NB, CNB, and RWB speech samples

A listening test was done for comparing performances between the proposed and conventional methods (Bhatt & Kosta, 2015; Chen & Leung, 2005, 2007; Chen et al., 2013; Prasad & Kishore Kumar, 2017; Sunil Kumar & Kishore Kumar, 2019). Here, WB signal, NB signal, CNB signal, and RWB signal were labelled I, II, III, and IV respectively. Participants are asked to do a pairwise comparison between the samples to tell whether the first sample was superior to, inferior than, or equal to the second. The responses after comparing I, II, and III with the other signals respectively are tabulated in Table 3, 4, 5. The number of participants with a specific preference is indicated by Arabic numerals in the table. It is observed that WB signal is superior to NB and CNB signals of traditional and the proposed methods from Table 3. Also, we observe that RWB signal quality is far

Table 3 Subjective comparison test results between I, II, III, and IV

Method	I	II	III	IV
Data hiding (Chen & Leung, 2005)	▷	30	30	14
	◁		0	0
	≈	0	0	16
Phonetic classification (Chen & Leung, 2007)	▷	30	30	12
	◁	0	0	0
	≈	0	0	18
Audio watermark (Chen et al., 2013)	▷	30	30	11
	◁	0	0	0
	≈	0	0	19
Side information (Bhatt & Kosta, 2015)	▷	30	30	9
	◁	0	0	0
	≈	0	0	21
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	▷	30	30	7
	◁	0	0	0
	≈	0	0	23
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	▷	30	30	6
	◁	0	0	0
	≈	0	0	24
Proposed method				
	▷	30	30	2
	◁	0	0	0
	≈	0	0	28

Table 4 Subjective comparison test results between II, III, and IV

	II	III	IV
Data hiding (Chen & Leung, 2005)	▷	8	3
	◁	4	18
	≈	18	9
Phonetic classification (Chen & Leung, 2007)	▷	8	1
	◁	2	19
	≈	20	10
Audio watermark (Chen et al., 2013)	▷	5	2
	◁	3	20
	≈	22	8
Side information (Bhatt & Kosta, 2015)	▷	5	2
	◁	2	22
	≈	23	6
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	▷	4	1
	◁	2	23
	≈	24	6
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	▷	3	1
	◁	2	24
	≈	25	5
Proposed method			
	▷	2	0
	◁	0	27
	≈	28	3

Table 5 Subjective comparison test results between III and IV

	III	IV
Data hiding (Chen & Leung, 2005)	▷	6
	◁	18
	≈	6
Phonetic classification (Chen & Leung, 2007)	▷	5
	◁	17
	≈	8
Audio watermark (Chen et al., 2013)	▷	3
	◁	18
	≈	9
Side information (Bhatt & Kosta, 2015)	▷	4
	◁	20
	≈	6
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	▷	2
	◁	23
	≈	5
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	▷	2
	◁	24
	≈	4
Proposed method		
	▷	0
	◁	29
	≈	1

superior using the proposed method over traditional methods from Table 3. Thus, the speech quality was enhanced by the proposed technique. Compared to traditional methods, it is observed that the RWB signal of the proposed method is superior to that of NB signal, as may be seen from Table 4. Also, a clear perceptual transparency improvement of the proposed method over the conventional methods was observed from Table 4 which shows that the quality of CNB signal is almost identical to that of NB signal. Obviously, the data embedding performed in the proposed method has very little impact on perception. Compared to conventional methods, it is observed that RWB speech of proposed technique is better than CNB speech from Table 5. Thus, the proposed method is demonstrated to produce a much better-quality speech signal than the conventional methods.

3.2 Objective quality evaluations

The database which was used in subjective listening tests was also used in evaluating objective measures. The perceptual transparency was assessed with the narrowband-perceptual evaluation of speech quality (NB-PESQ) measure (ITU-T, 2001). RWB speech quality was evaluated with the Log Spectral Distortion (LSD) (Chen & Leung, 2005, 2007) and wideband-perceptual evaluation of speech quality (WB-PESQ) measures (ITU-T, 2005). The robustness of hidden data against quantization and channel noises was evaluated with the help of a mean square error (MSE) measure (Chen & Leung, 2007).

3.2.1 RWB speech quality

The quality of RWB speech is evaluated using LSD measure and is calculated using the formula

$$LSD = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(20log_{10}\frac{\mathbf{g}_p}{a_s\left(e^{jw}\right)} - 20log_{10}\frac{\hat{g}_p}{\left|\hat{a}_s\left(e^{jw}\right)\right|}\right)^2 dw \tag{10}$$

where $\mathbf{g}_p$ is the gain of UB signal,$\frac{1}{a_s\left(e^{jw}\right)}$ is the SPENV of UB signal, $\hat{g}_p$ is the gain of the reconstructed UB signal and $\frac{1}{\hat{a}_s\left(e^{jw}\right)}$ is the SPENV of the reconstructed UB signal respectively. An RWB signal with the least value of LSD is said to be of good quality. The resultant LSD for conventional (Bhatt & Kosta, 2015; Chen & Leung, 2005, 2007; Chen et al., 2013; Prasad & Kishore Kumar, 2017; Sunil Kumar & Kishore Kumar, 2019) and proposed techniques with a μ-law channel model are presented in Table 6 and it was very evident that the RWB signal quality of the proposed technique was far superior to the signal quality generated using conventional techniques. In addition, the proposed technique offers an LSD of 2.2248 indicating that RWB speech of the proposed technique and original WB speech qualities are almost equal. Good RWB signal performance of the proposed technique which was already found in the subjective tests is now supported by these LSD values also. The proposed technique offers an LSD of 2.35 with the AWGN channel model.

Table 6 LSD test results

Method	Value
Data hiding (Chen & Leung, 2005)	12.83
Phonetic classification (Chen & Leung, 2007)	10.69
Audio watermark (Chen et al., 2013)	6.07
Side information (Bhatt & Kosta, 2015)	5.94
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	4.45
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	3.83
Proposed method	2.2248

Table 7 NB-PESQ test results

Method	Value
Data hiding (Chen & Leung, 2005)	2.87
Phonetic classification (Chen & Leung, 2007)	3.07
Audio watermark (Chen et al., 2013)	3.42
Side information (Bhatt & Kosta, 2015)	3.45
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	3.78
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	3.89
Proposed method	4.47

3.2.2 Perceptual transparency

The evaluation of perceptual transparency is done by providing NB and CNB signals as inputs and comparing them to rate speech quality. The NB-PESQ value will range between 0.5 and 4.5, where the higher the value, the more superior the quality. The average NB-PESQ values of conventional (Bhatt & Kosta, 2015; Chen & Leung, 2005, 2007; Chen et al., 2013; Prasad & Kishore Kumar, 2017; Sunil Kumar & Kishore Kumar, 2019) and proposed methods are tabulated in Table 7. The proposed technique gave NB-PESQ value of 4.47 which indicates that the proposed technique has excellent perceptual transparency over traditional techniques, which was already confirmed by subjective listening tests.

In Fig. 3, the upper plot a depicts the spectrogram of NB speech $y_{nb}(n)$, whereas the lower plot b depicts the spectrogram of the CNB signal $y^{1}_{nb}(n)$. It is clear from the figures that $y_{nb}(n)$ and $y^{1}_{nb}(n)$ are almost indistinguishable.

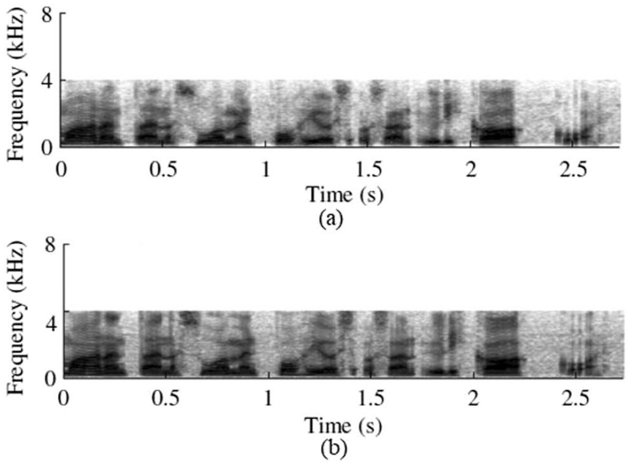

Fig. 3 Spectrograms from top to bottom: **a** Composite NB speech, **b** NB speech

3.2.3 Robustness of embedded information

AWGN with SNR ranges between 15 and 35 dB (Keiser and Strange 2017) is added to the CNB signal. The evaluation of the robustness of the proposed technique is done by utilizing MSE and is calculated using the formula.

$$MSE = \frac{1}{N}\sum_{n=0}^{N-1}\left(S^{1}_{wb}(n) - S_{wb}(n)\right)^{2} \tag{11}$$

where the RWB signal is represented by $S^{1}_{wb}(n)$ and the original WB signal is represented by $S_{wb}(n)$. The spreading sequence length is 16. An RWB signal with a small value of MSE is said to be of good quality. The proposed technique gave MSE values, as a function of SNR ranges between 15 and 35 dB, which are below 7.7083×10^{-4} indicating that the RWB signal quality obtained by the proposed technique is excellent. The proposed technique gave an MSE value after adding quantization noise (μ-law) to $\mathrm{s}^{1}_{\mathrm{nb}}(\mathrm{n})$ is 5.78×10^{-4} which indicates RWB signal quality that was obtained by the proposed technique is excellent.

3.2.4 WB speech quality

The evaluation of the quality of RWB speech is done by giving WB and RWB signals as inputs and comparing them in order to rate speech quality. The average WB-PESQ values of the conventional (Bhatt & Kosta, 2015; Chen & Leung, 2005, 2007; Chen et al., 2013; Prasad & Kishore Kumar, 2017; Sunil Kumar & Kishore Kumar, 2019) and proposed methods are shown in Table 8. A WB-PESQ value of 4.45 confirms that the RWB signal quality that was obtained by the proposed technique is excellent compared to traditional techniques, which was already confirmed by subjective listening tests on a set of participants. Thus, the speech quality was improved by using the proposed technique.

Table 8 WB-PESQ test results

Method	Value
Data hiding (Chen & Leung, 2005)	2.49
Phonetic classification (Chen & Leung, 2007)	2.73
Audio watermark (Chen et al., 2013)	3.64
Side information (Bhatt & Kosta, 2015)	3.71
Magnitude spectrum (Prasad & Kishore Kumar, 2017)	3.82
Transform domain (Sunil Kumar & Kishore Kumar, 2019)	3.98
Proposed method	4.45

4 Conclusions

In this paper, SBE utilizing the DCTBDH technique for extending the bandwidth of the existing NB telephone networks has been proposed. The spread spectral envelope parameters of the UB signal are embedded within the DCT coefficients of the NB signal at the transmitter. The embedded information is extracted at the receiver end to reconstruct the WB signal of good quality.

The spread spectrum technique is employed to increase the robustness of the embedded UB signal to quantization and channel noises by spreading the spectral envelope parameters by multiplying them with spreading sequences and then adding them up together to provide the embedded information. The embedded information can be reliably recovered by using a correlator. The robustness of the proposed method to quantization and channel noises is confirmed by MSE test. The MOS and NB-PESQ test values obtained for the proposed method indicate that the method embeds the UB information more transparently compared to the conventional methods. The RWB signal quality was enhanced by the proposed technique over conventional techniques and it was evident through subjective listening, LSD and WB-PESQ tests. The proposed method is demonstrated to produce a much better-quality speech signal than the conventional techniques. Hence it is suitable for extending the bandwidth of the existing telephone networks without making changes to the telephone networks.

References

Abel, J., & Fingscheidt, T. (2017). A DNN Regression Approach to Speech Enhancement by Artificial Bandwidth Extension. In *Proceedings of IEEE workshop on applications of signal processing to audio and acoustics*, pp. 219–223.

Archit, G., Brendan, S., Yannis, A. & Thomas, C. W. (2019). Speech bandwidth extension with wavenet. In *Proceedings of IEEE* *workshop on applications of signal processing to audio and acoustics*, pp. 205–208.

Berthy, F., Zeyu, J., Jiaqi, S., & Adam, F. (2019). Learning bandwidth expansion using perceptually-motivated loss. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 606–610.

Bhatt, N., & Kosta, Y. (2015). A novel approach for artificial bandwidth extension of speech signals by LPC technique over proposed GSM FR NB coder using high band feature extraction and various extension of excitation methods. *International Journal of Speech Technology, 18*(1), 57–64.

Bong-Ki, L., Kyoungjin, N., Joon-Hyuk, C., Kihyun, Ch., & Eunmi, O. (2018). Sequential deep neural networks ensemble for speech bandwidth extension. *IEEE Access, 6*, 27039–27047.

Chen, S., & Leung, H. (2005). Artificial bandwidth extension of telephony speech by data hiding. *In Proceedings of International Symposium on Circuits and Systems (ISCAS)*, pp. 3151–3154.

Chen, S., Leung, H., & Ding, H. (2007). Telephony speech enhancement by data hiding. *IEEE Transactions on Instrumentation and Measurement, 56*(1), 63–74.

Chen, S., & Leung, H. (2007). Speech bandwidth extension by data hiding and phonetic classification. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 593–596.

Chen, Z., Zhao, C., Geng, G., & Yin, F. (2013). An audio watermark based speech bandwidth extension method. *EURASIP Journal Audio, Speech and Music Processing, 10*, 1–8.

Dinan, E. H., & Jabbari, E. H. (1998). Spreading codes for direct sequence CDMA and wideband CDMA cellular networks. *IEEE Communications Magazine, 36*(9), 48–54.

ETSI ES 201 108 V1.1.2 (2000). Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.

Garofalo, J. S., Lamel, L. F., & Fisher, W. M. (2013). *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,* National Institute of Standards and Technology (NIST).

Geiser, B., Jax, P., & Vary, P. (2005). Artificial bandwidth extension of speech supported by watermark-transmitted side information. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1497–1500.

Geiser, B., & Vary, P. (2007). 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)*, pp. 533–536.

Geiser, B., & Vary, P. (2013). Speech bandwidth extension based on in-band transmission of higher frequencies. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 7507–7511.

Goldsmith, A. (2006). *Wireless communications*. Cambridge University Press.

Hanzo, L. L., Somerville, F. C. A., & Woodard, J. P. (2001). *Voice compression and communications: Principles and applications for fixed and wireless channels*. Wiley.

Hassan, A., Hershey, J. E., & Saulnier, G. J. (1998). *Perspectives in spread spectrum*. Kluwer Academic Publishers.

ITU-T. (2001). ITU-T Rec. P.862: 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. (2005). Recommendation P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs.

Jax, P. (2002). *Enhancement of bandlimited speech signals: Algorithms and theoretical bounds*. Ph.D. dissertation, RWTH Aachen University, Aachen, Germany.

Jax, P., & Vary, P. (2002). 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)*, pp. 237–240.

Jax, P., & Vary, P. (2003). On artificial bandwidth extension of telephone speech. *Signal Processing, 83*(8), 1707–1719.

Jax, P., & Vary, P. (2006). Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding? *IEEE Communication Magazine, 44*(5), 106–111.

Johannes, A., & Tim, F. (2019). Sinusoidal-based lowband synthesis for artificial speech bandwidth extension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(4), 765–776.

Jonas, S., Friedrich, F., Markus, B., & Gerhard, S. (2019). Artificial bandwidth extension using a conditional generative adversarial network with discriminative training. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 7005–7009).

Kanhe, A., & Aghila, G. (2016). DCT based Audio Steganography in Voiced and Un-voiced Frames. In Proceedings of *International Conference of Information and Analytics,* pp. 1–4.

Keiser, B. E., & Strange, E. (1995). *Digital telephony and network integration*. Van Nostrand Reinhold.

Kosta, Y. (2016). Simulation and overall comparative evaluation of performance between different techniques for high band feature extraction based on artificial bandwidth extension of speech over proposed global system for mobile full rate narrow band coder. *International Journal of Speech Technology, 19*(4), 881–893.

Kyoungjin, N., & Joon-Hyuk, Ch. (2020). Deep neural network ensemble for reducing artificial noise in bandwidth extension. *Digital Signal Processing, 102*, 1–6.

Mathieu, L., & Felix, G. (2020). Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 801–805.

Nilsson, M., & Kleijn. W. B. (2001). Avoiding overestimation in bandwidth extension of telephony speech. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing* (ICASSP), pp. 869–872.

Pramod, B., Massimiliano, T., & Nicholas, E. (2019). Latent representation learning for artificial bandwidth extension using a conditional variational auto-encoder. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 7010–7014.

Prasad, N., & Kishore Kumar, T. (2016). Bandwidth extension of speech signals: A comprehensive review. *International Journal of Intelligent Systems and Applications, 8*(2), 45–52.

Prasad, N., & Kishore Kumar, T. (2017). Speech bandwidth extension aided by spectral magnitude data hiding. *Circuits, Systems, and Signal Processing, 36*(11), 4512–4540.

Sagi, A., & Malah, D. (2007). Bandwidth extension of telephone speech aided by data embedding. *EURASIP Journal on Advances in Signal Processing, 2007*, 37–52.

Sunil Kumar, K., & Kishore Kumar, T. (2019). Speech Bandwidth Extension Aided by Hybrid Model Transform Domain Data Hiding. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5.

Xiang, H., Chenglin, X., Nana, H., Lei, X., EngSiong, Ch., & Haizhou, L. (2020). Time-domain neural network approach for speech bandwidth extension. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 866–870.

Yingwue, W., Shenghui, Z., & Dan, Q., (2016). Using conditional restricted Boltzmann machines for spectral envelope modelling in speech bandwidth extension. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, pp. 5930–5934.

Yuanjie, D., Yaxing, L., Xiaoqi, L., Shan, X., Dan, W., Zhihui, Z., & Shengwu, X. (2020). A time-frequency network with channel attention and non-local modules for artificial bandwidth extension. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP),* pp. 6954–6958.

Zhen-Hua, L., Yang, A., & Yu, G. (2018). Waveform modelling and generation using hierarchical recurrent neural networks for speech bandwidth extension. *IEEE/ACM Transaction Audio, Speech, and Language Process, 26*(5), 883–894.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.