Artificial Bandwidth Extension of Narrowband Speech Signals for the Improvement of Perceptual Speech Communication Quality

Nam In Park, Young Han Lee, and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{naminpark,cpumaker,hongkook}@gist.ac.kr

Abstract. In this paper, an artificial bandwidth extension (ABE) algorithm from narrowband to wideband is proposed in order to improve the quality of narrowband speech. The proposed ABE algorithm is based on spectral band replication in the modified discrete cosine transform (MDCT) domain with no additional bits. In particular, the patch index search for the replication band is restricted so that the harmonic structure of the wideband speech is maintained after ABE. In the proposed ABE algorithm, we first determine whether the current analysis frame of speech signals is voiced or unvoiced. A harmonic spectral replication or a correlation-based replication approach is then applied for the voiced or unvoiced frame, respectively. The proposed ABE algorithm is finally embedded into the G.729 speech decoder as a post-processor. It is shown from the subjective evaluation using a MUSHRA test that the mean opinion score of the wideband speech signals extended by the proposed ABE method is measured as 75.5, which is higher of around 14% than that of narrowband speech signals.

Keywords: Bandwidth extension, artificial bandwidth extension, narrowband speech, wideband speech, perceptual quality improvement.

1 Introduction

Applications using wideband speech are slowly yet inevitably gaining over those using narrowband speech. Such a recent trend requires speech coding technology with an increased quality of decoded signals, instead of concentrating on the absolute compression efficiency [1-3]. In most speech communication systems, the speech bandwidth is limited to a range of 0.3–3.4 kHz. This speech bandwidth represents a good compromise between speech quality and transmission bandwidth for voiced sounds in general, but often a poor one for unvoiced sounds; this fact typically results in muffling on speech quality. In order to mitigate such a problem, wideband speech coding has been proven as an alternative [4]. Indeed, wideband speech, whose bandwidth ranges from 50 Hz to 7 kHz, spans all distinctive speech frequency components, thus it sounds clearer and gives a more natural conversation than that using narrowband speech.

T.-h. Kim et al. (Eds.): FGCN 2011, Part II, CCIS 266, pp. 143–153, 2011.

However, narrowband speech has been popularly serviced in many applications such as voice communications over a public switched telephone network (PSTN), voice over IP (VoIP), and voice applications in smart phones [5]. Therefore, simply replacing a narrowband speech codec with a wideband one in order to improve the quality of decoded speech is not a right solution. Instead, the extension of speech bandwidth from narrowband to wideband could be an alternative.

There are two different kinds of approaches for extending the bandwidth according to whether or not the side information is available. It is usual to realize bandwidth extension by using the side information that is transmitted from the encoder. In other words, the encoder should generate auxiliary information based on the analysis of the high frequency component of the input signal [6-7]. The decoder recovers the high frequency signal from the low frequency signal and then uses the auxiliary information to adjust the generated high frequency signal. For example, the G.729.1 speech coder provides embedded coding with 12 different bit rates between 8 and 32 kbit/s [7]. The baseline coder of G.729.1 is fully compatible with G.729, thereby ensuring narrowband speech quality in 8 kbit/s mode. From the 14 kbit/s mode, whose operation mode is called 'layer 2,' a wideband signal can be synthesized using a BWE technique. By allocating additional bits for the BWE, the high band signal can be reconstructed in the decoder. However, this BWE approach requires additional bits and also requires some modification of the encoding process of a speech coder.

On the other hand, instead of using the side information, the other type of BWE, which is also called artificial bandwidth extension (ABE), can estimate the high band signal from the low band signal without any side information. The estimation can be done by using a pattern recognition algorithm such as hidden Markov models (HMMs) [8], Gaussian mixture models (GMMs) [9], and so on [10-12]. For example, an ABE method is based on a source filter model of speech production, according to the fact that speech consists of an excitation signal and a vocal tract filter [13]. The vocal track filter is usually modeled by a set of linear prediction (LP) coefficients. Based on statistical recovery, GMM is then applied in order to extend the envelope. Such a method is realized without any additional bits, while it requires a training process.

Instead of a model-based approach for predicting the high band from the low band, another conventional method was the spectral band replication (SBR) based ABE. This method copies modified discrete cosine transform (MDCT) spectrum to generate the high-frequency signal, and then adjusts its tonality for improving the subjective quality [14]. However, this method can cause a mismatch of the harmonic component at the boundaries between low and high bands, since it simply copies the low band to construct the high band without considering harmonic structure of speech signal.

In this paper, an ABE method based on harmonic spectral replication and correlation-based spectral replication is proposed, where the extension is performed in the modified discrete cosine transform (MDCT) domain. Since the voice signal consists of tones and harmonics, applying a correlation-based method can cause a mismatch of the harmonic characteristics. Hence, we first classify each analysis frame of speech signal into a voiced or an unvoiced frame using a spectral tilt parameter, and then apply the harmonic spectral replication or correlation-based spectral replication

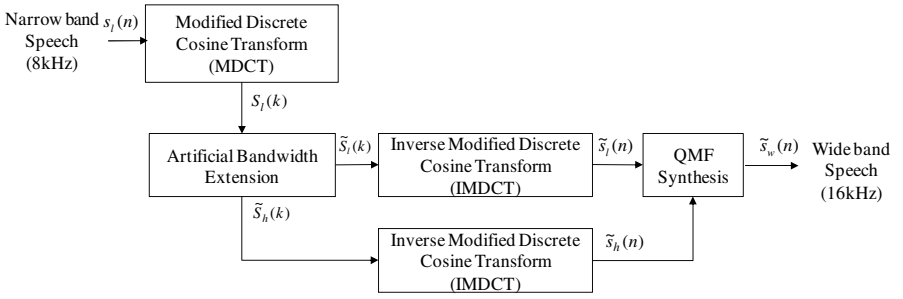

Fig. 1. Block General structure of artificial bandwidth extension from narrowband speech to wideband speech, which is applied in the MDCT domain

for the voiced or unvoiced frame, respectively. In other words, the harmonic spectral replication can maintain the harmonic characteristic between low and high band for voiced frames.

The remainder of this paper is organized as follows. Following this introduction, Section 2 describes the proposed ABE algorithm, and Section 3 describes how to realize the proposed ABE algorithm on a CELP-type speech decoder. Section 4 then demonstrates the performance of the proposed ABE algorithm. Finally, this paper is concluded in Section 5.

2 Proposed Artificial Bandwidth Extension (ABE) Algorithm

Fig. 1 shows a general structure of ABE applied in the MDCT domain to extend bandwidth from narrowband to wideband. In the figure, the narrowband speech, $s(n)$, is segmented into a sequence of frames, where frame size is *N*. Then, each analysis frame is transformed into the frequency domain using a 2*N*-point MDCT, $S_l(k)$. After that, an ABE algorithm is applied in order to obtain high band MDCT coefficients, $S_h(k)$. In this paper, the proposed ABE algorithm based on the combination of harmonic spectral replication and correlation-based spectral replication is applied, which will be explained in Section 2.1. In addition, in order to prevent MDCT coefficients in the boundary between narrowband and high band from being abruptly changed, the MDCT coefficients of narrowband and high band, $S_l(k)$ and $S_h(k)$ are modified into $\tilde{S}_l(k)$ and $\tilde{S}_h(k)$. Next, the low band and high band signals, $\tilde{s}_l(n)$ and $\tilde{s}_h(n)$, in the time domain are obtained by applying a 2*N*-point inverse MDCT (IMDCT), respectively. Finally, the bandwidth extended signal, $\tilde{s}_w(n)$, is obtained by the quadrature mirror filterbanks (QMF).

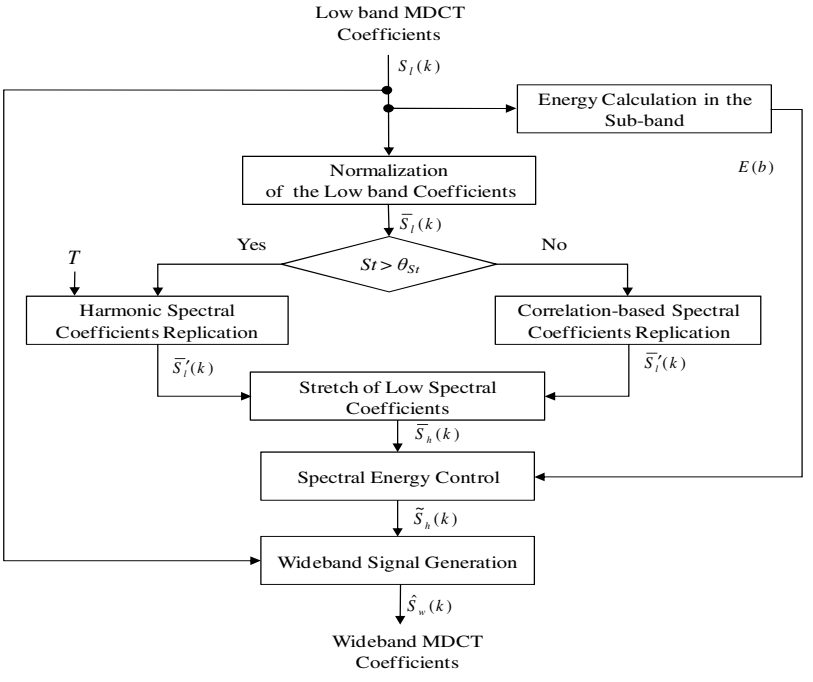

Fig. 2. Block diagram of the proposed ABE algorithm

2.1 U/V-Dependent Spectral Band Replication

Fig. 2 shows a block diagram of the proposed ABE algorithm. The proposed ABE algorithm generates the low and high band MDCT coefficients, $\tilde{S}_l(k)$ and $\tilde{S}_h(k)$. In this paper, the frame size, *N*, is set to 512. First, 2*N*-point MDCT coefficients of the low band, $S_l(k)$, are grouped into 16 sub-bands. That is, each sub-band has 32 MDCT coefficients. Then, the energy of the *b*-th sub-band, $E(b)$, is defined as

$$E(b) = \sqrt{\sum_{k=32\cdot b}^{k=32\cdot(b+1)-1} S_l^2(k)}, \qquad b = 0,1,\cdots,15 . \tag{1}$$

Next, $E(b)$ is used to normalize each MDCT coefficient belonging to the *b*-th sub-band, such as

$$\overline{S}_l(k) = \frac{S_l(k)}{E(b)}, \qquad 32b \le k < 32(b+1) \text{ and } b = 0,1,\cdots,15 \tag{2}$$

where $\overline{S}_l(k)$ is the *k*-th normalized low band MDCT coefficient.

In order to classify each frame into a voice or an unvoiced frame, we define spectral tilt, St, by using the first reflection coefficient as

$$St = \frac{\sum_{n=1}^{N-1} s(n)s(n-1)}{\sum_{n=1}^{N-1} s^2(n)} \tag{3}$$

where $s(n)$ is the narrowband speech signal at the n-th time instant as shown in Fig. 1. If St is greater than θ_{St}, this frame is declared a voiced frame; otherwise, it is as an unvoiced frame. In this paper, θ_{St} is set to 0.25 from the preliminary experiment. Next, in order to extract the pitch information for a voiced frame, a normalized autocorrelation function is calculated as

$$R(\tau) = \frac{\sum_{n=\tau}^{N-1} s(n)s(n-\tau)}{\sqrt{\sum_{n=\tau}^{N-1} s^2(n)}} \quad . \tag{4}$$

Using Eq. (4), the pitch is obtained by selecting τ at which $R(\tau)$ is maximized. That is, $T = \arg\max_{P_l \le \tau \le P_h} R(\tau)$, where P_l and P_h are set to 20 and 147, respectively [15]. Specifically, in order to generate a high band signal with the harmonic characteristics, the harmonic period in the MDCT domain is determined as

$$\Delta_v = 2N/T \ . \tag{5}$$

The k-th harmonic MDCT coefficient, $\overline{S}_l'(k)$, is then expressed as

$$\overline{S}_l'(k) = \overline{S}_l(k + \frac{N}{2} - \lfloor \Delta_v - \mathrm{mod}(N, \Delta_v) \rfloor), \ k = 0,1,\cdots,\frac{N}{2} - 1 \tag{6}$$

where $\overline{S}_l(k)$ is the normalized low band signal as described in Eq. (2), and $\mathrm{mod}(x, y)$ is the modulus operation defined as $\mathrm{mod}(x, y) = x \,\%\, y$. Also $\lfloor x \rfloor$ is the largest integer less than or equal to x.

On the other hand, in order to patch high band MDCT coefficients from low band MDCT coefficients for an unvoiced frame, the optimal position in $\overline{S}_l(k)$ is determined by the equation of

$$\Delta_{uv} = \underset{0 \le m \le N/4-1}{\mathrm{argmax}} \ [corr(S_l(k), \overline{S}_l(k+m))] \tag{7}$$

where Δ_{uv} is the optimum shift of the patch and $corr(S_l(k), \overline{S}_l(k+m))$ is the cross-correlation between the low band MDCT coefficients, $S_l(k)$, and the normalized low band MDCT coefficients, $\overline{S}_l(k)$. In other words, the cross-correlation can be represented as

$$corr(S_l(k), \overline{S}_l(k+m)) = \sum_{k=0}^{N/4-1} S_l(k+\frac{3}{4}N)\,\overline{S}_l(k+m), \quad m = 0,1,\cdots,\frac{N}{4}-1. \tag{8}$$

Finally, $\overline{S}'(k)$ that is the most correlated to $\overline{S}_l(k)$ in a range of 3–4 kHz is expressed as

$$\overline{S}_l'(k) = \overline{S}_l(k+\frac{1}{4}N+\Delta_{uv}), \quad k = 0,1,\cdots,\frac{N}{2}-1. \tag{9}$$

The MDCT coefficients obtained from Eqs. (6) or (9), $\overline{S}_l'(k)$, which have a 2 kHz bandwidth, are finally stretched to the extended MDCT coefficients with a 4 kHz bandwidth by the following equation of

$$\overline{S}_h(k) = \begin{cases} \overline{S}_l'(k/2), & k = 0,2,\cdots,N-2 \\ 0, & k = 1,3,\cdots,N-1 \end{cases} \tag{10}$$

where $\overline{S}_h(k)$ is the k-th normalized and extended MDCT coefficient.

2.2 Energy Control

In order to avoid an abrupt change in energy at the high band after patching MDCT coefficients from the low band, the amplitude of each MDCT coefficient for the high band should be adjusted. First of all, the refined energy for the b-th high band, $E_h(b)$, is defined as

$$E_h(b) = \begin{cases} \alpha \cdot E(b+7), & if\ E(b+8) > \alpha \cdot E(b+7) \\ E(b+8), & otherwise \end{cases},\ b = 0,1,\cdots,7 \tag{11}$$

where $E(b)$ is the energy in the b-th low band as defined in Eq. (1) and α is set to 1.1 in this paper. Next, in order to maintain the energy of boundary between the low and high band, the scale factor of the energy, β, is calculated as

$$\beta = E(15) \Big/ E_h(0) \tag{12}$$

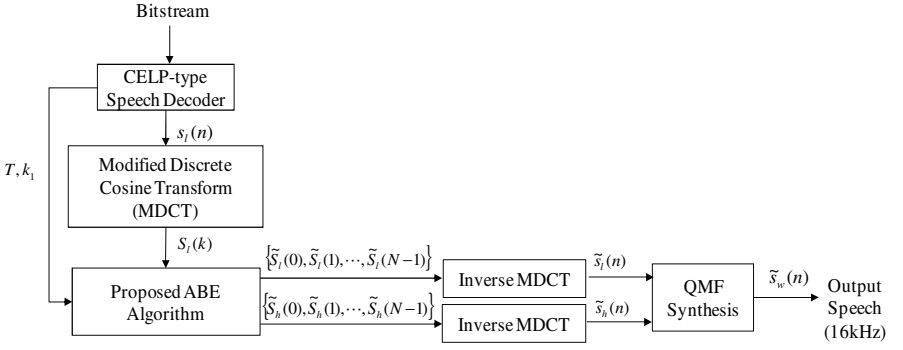

Fig. 3. Application of the proposed ABE algorithm as a post-processor of a CELP-type narrowband speech decoder

where $E(15)$ is the energy of the last sub-band in the low band and $E_h(0)$ is the energy of the first sub-band in the high band. By combining Eqs. (10) and (11), the energy of the *b*-th sub-band in the high band, $\hat{E}_h(b)$, is finally modified as

$$\hat{E}_h(b) = \beta \cdot E_h(b), \quad b = 0,1,\cdots,7 \ . \tag{13}$$

Similarly to Eq. (10), $\hat{E}_h(b)$ is stretched as

$$\overline{E}_h(b) = \begin{cases} \hat{E}_h(b/2), & b = 0,2,\cdots,14 \\ \overline{E}_h(b-1), & b = 1,3,\cdots,15 \end{cases} \ . \tag{14}$$

Next, the amplitude of $\tilde{S}_h(k)$ is adjusted as

$$\tilde{S}_h(k) = \overline{S}_h(k)\overline{E}_h(b), \quad b = \lfloor k/32 \rfloor, \quad k = 0,1,\cdots,N-1 \ . \tag{15}$$

Then, the wideband MDCT coefficients, $\tilde{S}_w(k)$, are constructed by integrating $\tilde{S}_l(k)$ and $\tilde{S}_h(k)$, such as

$$\tilde{S}_w(k) = [\tilde{S}_l(0), \tilde{S}_l(1), \cdots, \tilde{S}_l(N-1), \tilde{S}_h(0), \tilde{S}_h(1), \cdots, \tilde{S}_h(N-1)] \ . \tag{16}$$

3 Application of the Proposed ABE Algorithm as a Post-processor of a CELP-Type Narrowband Speech Coder

The proposed ABE algorithm is applied to the reconstructed speech by the G.729 speech decoder [16] which is one of CELP-type narrowband speech coders. Fig. 3

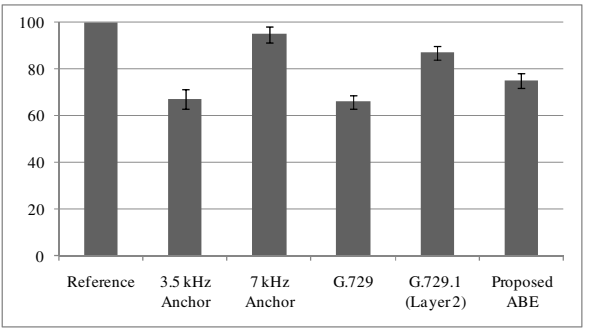

Fig. 4. MUSHRA test results

shows how the proposed ABE algorithm works as a post-processor of the decoder. For a given bitstream, the G.729 decoder reconstructs speech signals, $s_l(n)$. Note here that G.729 encodes narrowband speech signals once every 10 msec and compresses them with a bit-rate of 8 kbit/s, thus the reconstructed speech signals of 10 msec long are processed by the proposed ABE algorithm. That is, the frame size, *N*, is equal to 80, and $s_l(n)$ is transformed into the frequency components using a 160-point MDCT.

In order to realize the proposed ABE algorithm, we require the spectral tilt parameter denoted in Eq. (3). Instead of directly computing the spectral tilt parameter from $s_l(n)$, the first reflection coefficient, k_1, which can be obtained from the spectral envelope parameters during G.729 decoding, is used for St in Eq. (3). In addition, when St is greater than θ_{St}, a pitch period should be computed. Here, the pitch information obtained from the decoding is also used for T in Eq. (5).

After applying the proposed ABE algorithm, the low band and high band signals, $\tilde{s}_l(n)$ and $\tilde{s}_h(n)$, are obtained by a 160-point IMDCT, respectively. Finally, the bandwidth extended signal, $\tilde{s}_w(n)$, is synthesized by filtering the 64-QMF [7].

4 Performance Evaluation

In order to demonstrate the effectiveness of the proposed ABE algorithm, a multiple stimuli with hidden reference and anchor (MUSHRA) listening test [17] and a spectrum comparison were carried out.

For the MUSHRA test, 6 speech sentences, comprised of the utterances of 3 males and 3 females, were taken from the sound quality assessment material (SQAM) [18]. Especially, since SQAM speech files were recorded with stereo at a sampling rate of 44.1 kHz, the right channel signals of each file were down-sampled from 44.1 kHz to two different versions such as 8 and 16 kHz. In other words, 8 kHz down-sampled

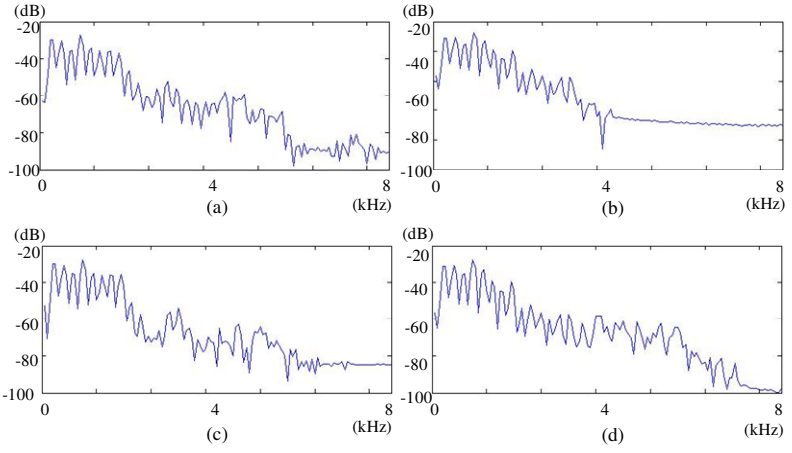

Fig. 5. Comparison of the spectra obtained from (a) the original speech signals sampled at 16 kHz, (b) the decoded signals by G.729 decoder, (c) the decoded signals by G.729.1(Layer 2), and (d) the speech signal by the proposed ABE algorithm

signals were processed by G.729 and further by the proposed ABE, while 16 kHz down-sampled signals were processed for G.729.1(Layer 2). Next, two anchors with cut-off frequencies of 3.5 and 7 kHz were prepared for the MUSHRA test.

Seven people with no auditory disease participated in this test. The listener was presented with several audio stimuli. The first was the reference, which was the original speech down-sampled at 16 kHz. The remainders were processed by G.729, G.729.1(Layer 2) and the proposed ABE method. Each listener gave a score between 0 and 100 depending upon their opinion of the quality.

Fig. 4 shows the results of the MUSHRA test. As shown in the figure, the proposed method gave an average score of 75.5, which was higher than G.729 and lower than G.729.1(Layer 2), as we expected. This result indicated that the proposed ABE method enhanced the performance of G.729 without any additional bits. Moreover, compared to G.729, the quality improvement of 43% was achieved by the proposed ABE method.

Finally, the spectra of speech signals processed by G.729, G.729.1(Layer 2), and the proposed method were compared, which is shown in Fig. 5. It was shown from the figure that the high band spectrum of the proposed ABE algorithm was quite similar to that of G.729.1(Layer 2).

5 Conclusion

In this paper, we proposed an artificial bandwidth extension (ABE) algorithm from narrowband to wideband to improve the quality of narrowband speech. The proposed ABE algorithm was based on the harmonic spectral replication and correlation-based spectral replication for voice and unvoiced frame, respectively. The proposed ABE

algorithm was embedded into the G.729 speech decoder as a post-processor. It was shown from the subjective evaluation that the proposed ABE method achieved MUSHRA score improvement of 43% compared to G.729. Moreover, when comparing the spectra of speech signal, it was shown that the high band spectrum of speech signals processed by the proposed ABE algorithm was quite similar to that by the layer 2 mode of G.729.1 that was one of standardized wideband speech coders in ITU-T.

Acknowledgments. This work was supported in part by the Practical R&D Program of the GIST Technology Initiative (GTI), Gwangju Institute of Science and Technology, Korea.

References

1. Mikko, T., Lasse, L., Anssi, R., Henri, T.: Scalable super-wideband extension for wideband coding. In: Proceedings of ICASSP, pp. 161–164 (2009)
2. Stephen, V.: Listener ratings of speech passbands. In: Proceedings of IEEE Workshop on Speech Coding, pp. 81–82 (1997)
3. Park, N.I., Kim, H.K., Jung, M.A., Lee, S.R., Choi, S.H.: Burst packet loss concealment using multiple Codebooks and comfort noise for CELP-type speech coders in wireless sensor networks. Sensors 11(5), 5323–5336 (2011)
4. ITU-T Recommendation G.830: Subjective Performance Assessment of Telephone-band and Wideband Digital Codec (1996)
5. Goode, B.: Voice over internet protocol (VoIP). Proceedings of the IEEE 90(9), 1495–1517 (2002)
6. Kosuke, T., Kei, K.: Low-complexity bandwidth extension in MDCT domain for low-bitrate speech coding. In: Proceedings of ICASSP, pp. 4145–4148 (2009)
7. Rogot, S., Kovesi, B., Trilling, R., Virette, D., Duc, N., Massaloux, D., Proust, S., Geiser, B., Gartner, M., Schandl, S., Taddei, H., Yang, G., Shlomot, E., Ehara, H., Yoshida, K., Vaillancourt, T., Salami, R., Lee, M.S., Kim, D.Y.: ITU-T G.729.1: an 8-32 kbit/s scalable coder interoperable with G.729 for wideband Telephony and voice over IP. In: Proceedings of ICASSP, pp. 529–532 (2007)
8. Jax, P., Vary, P.: On artificial bandwidth extension of telephone speech. Signal Processing 83, 1707–1719 (2003)
9. Song, G.-B., Martynovich, P.: A study of HMM-based bandwidth extension of speech signals. Signal Processing 89, 2036–2044 (2009)
10. Kornagel, U.: Techniques for artificial bandwidth extension of telephone speech. Signal Processing 86, 1296–1306 (2006)
11. Pulakka, H., Laaksonen, L., Vainio, M., Pohjalainen, J., Alku, P.: Evaluation of an artificial speech bandwidth extension method in three languages. IEEE Transactions on Audio, Speech, and Language Processing 16, 1124–1137 (2008)
12. Kim, K., Lee, M., Kang, H.: Speech bandwidth extension using temporal envelope modeling. IEEE Signal Processing Letters 15, 429–432 (2008)
13. Murali, M.D., Karpur, D.B., Narayan, M., Kishore, J.: Artificial bandwidth extension of narrowband speech using Gaussian mixture model. In: Proceedings of ICASSP, pp. 410–412 (2011)

14. Tsujino, K., Kikuiri, K.: Low-complexity bandwidth extension in MDCT domain for low-bitrate speech coding. In: Proceedings of ICASSP, pp. 4145–4148 (2009)
15. Kondoz, A.M.: Digital Speech: Coding for Low Bit Rate Communication Systems, 2nd edn. Weiley (2004)
16. ITU-T Recommendation G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Code-Excited Linear Prediction, CS-ACELP (1996)
17. ITU/ITU-R BS 1534: Method for Subjective Assessment of Intermediate Quality Level of Coding Systems (2001)
18. EBU.: Sound Quality Assessment Material Recording for Subjective Tests (1988)