SPEECH BANDWIDTH EXTENSION BASED ON IN-BAND TRANSMISSION OF HIGHER FREQUENCIES

Bernd Geiser and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{geiser|vary}@ind.rwth-aachen.de

ABSTRACT

A new method for wideband speech transmission is proposed which is fully backwards compatible with narrowband telephone systems. For this purpose, a pitch-scaled version of the higher speech frequencies (4 – 6.4 kHz) is inserted into the previously "unused" 3.4 – 4 kHz frequency range of standard telephone speech. This operation is reverted at the decoder side. A consistently good wideband speech quality can be achieved, even after transmission over common codecs and codec tandems. The quality impact on the narrowband part of the signal is insignificant.

***Index Terms*—** wideband speech transmission, speech bandwidth extension, pitch scaling

1. INTRODUCTION

Wideband speech transmission with a higher audio bandwidth than the traditional 0.3 – 3.4 kHz frequency band is an essential feature for contemporary high-quality speech communication systems. Suitable codecs, such as the AMR-WB [1, 2], are available and offer a significantly increased speech quality and intelligibility compared to narrowband telephony. However, the requirement of *backwards compatibility* with existing equipment effectively precluded a timely deployment of the new technology. For example, "HD-Voice" transmission in cellular networks is only slowly being introduced.

Moreover, even if wideband transmission is supported by the receiving terminal and by the corresponding network operator, still the *calling* terminal or parts of the involved transmission chain may employ only narrowband codecs. Therefore, subscribers of HD-voice services will still experience inferior speech quality in many cases.

1.1. Relation to Prior Work

This paper presents a new solution for a fully *backwards compatible* transmission of wideband speech signals. In the literature, several attempts to maintain such compatibility have appeared, first to name techniques for "artificial bandwidth extension" (ABWE) of speech, i.e., (statistical) estimation of missing frequency components from the narrowband signal alone, e.g., [3, 4, 5]. For ABWE, there are in fact no further prerequisites apart from the mere availability of narrowband speech. Although this "receiver-only" approach constitutes the most generic solution, it suffers from an inherently limited performance which is not sufficient for the regeneration of *high quality* wideband speech signals.

A much better wideband speech quality is obtained when some compact side information about the upper frequency band is explicitly transmitted, e.g., [6, 7]. In this case, the backwards compatibility w.r.t. the transmission network can be maintained with steganographic methods that hide the side information bits in the narrowband signal or in the respective bitstream by using signal-domain watermarking techniques, e.g., [8, 9], or "in-codec" steganography, e.g., [10, 11, 12]. The signal domain watermarking approach is, however, not robust against low-rate narrowband speech coding and, in practice, requires tedious synchronization and equalization procedures. The "in-codec" techniques, in contrast, facilitate relatively high hidden bit rates, but, owing to the strong dependence on the specific speech codec, any hidden information will be lost in case of transcoding.

1.2. Proposed Transmission System

The proposed transmission system constitutes an alternative to previous, steganography-based methods for backwards compatible wideband communication. The basic idea is to insert a pitch-scaled version of the higher frequencies (4 – 6.4 kHz in this paper) into the previously "unused" 3.4 – 4 kHz frequency range of standard telephone speech which corresponds to a down-scaling factor of $\rho = (4 - 3.4)/(6.4 - 4) = \frac{1}{4}$. This operation is reverted at the decoder side (up-scaling factor $1/\rho = 4$).

Of the numerous pitch-scaling methods which are available, cf. [13], a comparatively simple DFT-domain technique turned out to be well-suited for our purposes, because, in this case, the pitch scaling and the required frequency domain insertion/extraction operations can be carried out within the same signal processing framework. Besides, the concerned higher speech frequencies do not contain any dominant tonal components that could be problematic for the pitch scaling algorithm.

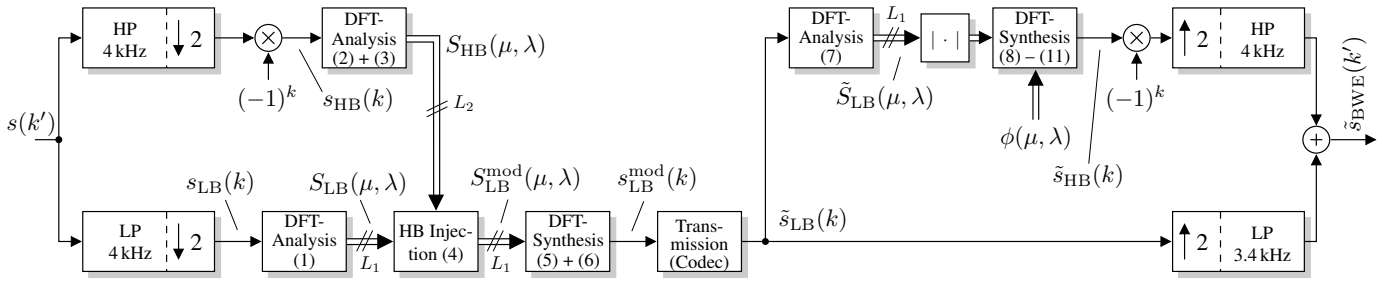

Fig. 1. System overview (bracketed numbers reference the respective equations in the text).

2. ENCODER

At the encoder side of the proposed system, shown in the left part of Figure 1, the wideband speech signal $s(k')$ with its sampling rate of $f_{\mathrm{s}}' = 16\,\mathrm{kHz}$ is first analyzed. Then the high frequency analysis result is inserted into the lower band. Finally, the modified narrowband speech $s_{\mathrm{LB}}^{\mathrm{mod}}(k)$ is synthesized. The sampling rate of the subband signals is $f_{\mathrm{s}} = 8\,\mathrm{kHz}$.

2.1. Analysis of Wideband Speech

The wideband signal $s(k')$ is first split into the two subband signals $s_{\mathrm{LB}}(k)$ and $s_{\mathrm{HB}}(k)$, e.g., with a half-band QMF filterbank. Then, for the lower frequency band in frame λ, a windowed DFT analysis is performed using a long window length L_1 and a large window shift S_1:

$$S_{\mathrm{LB}}(\mu,\lambda) = \sum_{k=0}^{L_1-1} s_{\mathrm{LB}}(k+\lambda S_1) w_{L_1}(k) \cdot \mathrm{e}^{-2\pi j \frac{k\mu}{L_1}} \qquad (1)$$

for $\mu \in \{0,\ldots,L_1-1\}$. The window function $w_{L_1}(k)$ is the square root of a Hann window of length L_1. We have chosen $L_1 = 128$ and $S_1 = 32$ which yields a temporal resolution of $S_1/f_{\mathrm{s}} = 4\,\mathrm{ms}$. The high band is analyzed with the same (large) window shift S_1, but with less spectral resolution, i.e., with a *shorter* window of length $L_2 = \rho \cdot L_1 = 32$:

$$S_{\mathrm{HB}}(\mu,\lambda) = \sum_{k=0}^{L_2-1} s_{\mathrm{HB}}(k+\kappa(\lambda)+\lambda S_1) w_{L_2}(k) \cdot \mathrm{e}^{-2\pi j \frac{k\mu}{L_2}} \qquad (2)$$

for $\mu \in \{0,\ldots,L_2-1\}$. Thereby, the actual window shift for frame λ is modified by the term $\kappa(\lambda)$ which is given as

$$\kappa(\lambda) = \arg \min_{\kappa \in \{-\kappa_0,\ldots,\kappa_0\}} \sum_{k=0}^{L_2-1} s_{\mathrm{HB}}^2(k+\kappa+\lambda S_1) \qquad (3)$$

with $\kappa_0 = 8$. This energy-minimizing choice of the window shift avoids audible fluctuations in the overall output signal $\tilde{s}_{\mathrm{BWE}}(k')$. Note that the sequence of analysis windows in (2) does not necessarily overlap which, in effect, realizes the time-stretching by a factor of $1/\rho$ (or, respectively, the pitch-scaling by a factor of ρ).

2.2. High Frequency Injection

The analysis procedure, as described above, has been designed such that $(4\,\mathrm{kHz} - 3.4\,\mathrm{kHz}) \cdot L_1 \stackrel{!}{=} 2.4\,\mathrm{kHz} \cdot L_2$, i.e., the first 2.4 kHz of the analysis result of (2) fit in the upper 600 Hz of the analysis result of (1). Omitting the frame index λ as well as the (implicit) complex conjugate symmetric extension for $\mu > \frac{L_1}{2}$, the high band injection procedure for the signal magnitude can be written as:

$$\left|S_{\mathrm{LB}}^{\mathrm{mod}}(\mu)\right| = \begin{cases} |S_{\mathrm{LB}}(\mu)| & \text{for } \mu < \mu_0 \\ g_{\mathrm{e}} \frac{L_1}{L_2} \cdot |S_{\mathrm{HB}}(\mu-\mu_0)| & \text{for } \mu_0 \le \mu \le \mu_1 \end{cases} \qquad (4)$$

with $\mu_0 = \frac{L_1 - \lceil 2.4/4 \cdot L_2 \rceil}{2}$ and $\mu_1 = \frac{L_1}{2}$. With (4), the upper 600 Hz of $|S_{\mathrm{LB}}(\mu)|$ are overwritten with the high band magnitude spectrum. The "injection gain" g_{e} has been set to 1 in our experiments; higher values for g_{e} can improve the robustness of the injected high band information against channel or coding noise, if desired. Note that the *phase* of $S_{\mathrm{LB}}(\mu)$ is not modified here. Nevertheless, it can also be included in (4) to facilitate different high band reconstruction mechanisms, cf. Section 3.2.

2.3. Narrowband Resynthesis

The composite signal $S_{\mathrm{LB}}^{\mathrm{mod}}(\mu)$ is now transformed into the time domain by reverting the lower band analysis of (1), i.e., the IDFT uses the longer window length of L_1:

$$s_{\mathrm{LB}}^{\mathrm{mod}}(k,\lambda) = \frac{1}{L_1} \sum_{\mu=0}^{L_1-1} S_{\mathrm{LB}}^{\mathrm{mod}}(\mu,\lambda) \cdot \mathrm{e}^{2\pi j \frac{k\mu}{L_1}} \qquad (5)$$

for $k \in \{0,\ldots,L_1-1\}$ and 0 outside the frame interval. The subsequent overlap-add procedure uses the larger window shift S_1, i.e.:

$$s_{\mathrm{LB}}^{\mathrm{mod}}(k) = \sum_{\lambda} s_{\mathrm{LB}}^{\mathrm{mod}}(k-\lambda S_1,\lambda)\, w_{L_1}(k-\lambda S_1) \qquad (6)$$

for all k. Note that, for compatibility reasons, the speech quality of $s_{\mathrm{LB}}^{\mathrm{mod}}(k)$ must not be degraded compared to the original narrowband speech $s_{\mathrm{LB}}(k)$. This is examined in Section 4.1. Example spectrograms of $s_{\mathrm{LB}}^{\mathrm{mod}}(k)$ and, for comparison, $s_{\mathrm{LB}}(k)$ are shown in left part of Figure 2.

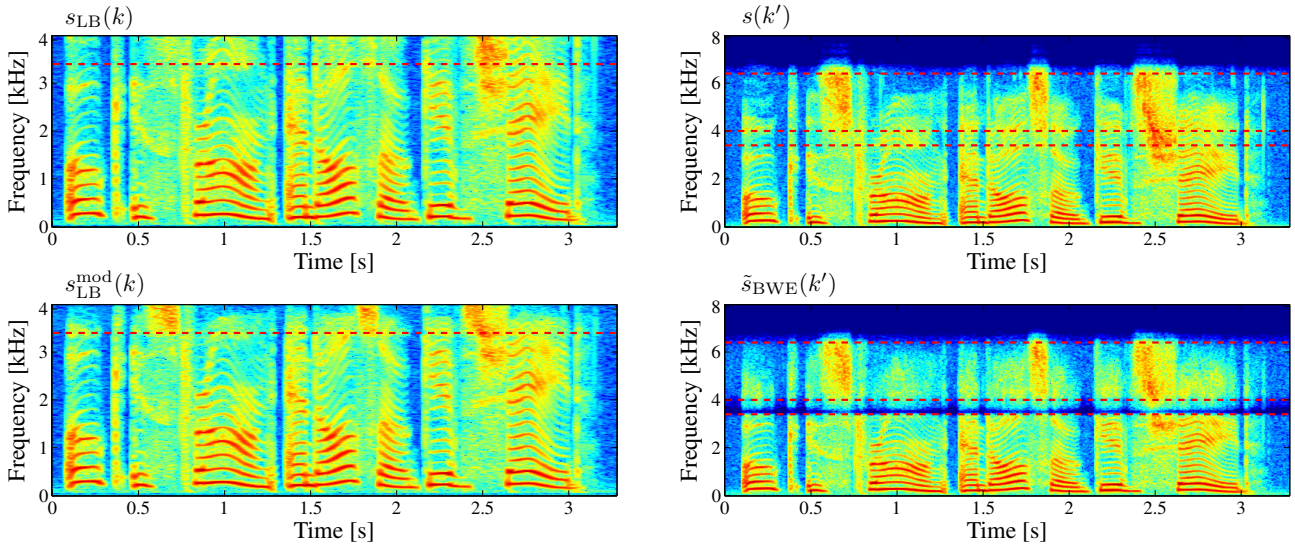

Fig. 2. Spectrograms for an exemplary input speech signal (red lines are placed at 3.4, 4, and 6.4 kHz).

3. DECODER

At the decoder side, shown in the right part of Figure 1, the received narrowband signal, denoted $\tilde{s}_{\mathrm{LB}}(k)$, is first analyzed, then the contained high band information is extracted and a high band signal $\tilde{s}_{\mathrm{HB}}(k)$ is synthesized which is finally combined with the narrowband signal to form the bandwidth extended output signal $\tilde{s}_{\mathrm{BWE}}(k')$.

3.1. Analysis of the Received Narrowband Signal

The decoder side analysis of $\tilde{s}_{\mathrm{LB}}(k)$ uses the long window length L_1, but a *small* window shift $S_2 = \rho \cdot S_1 = 8$:

$$\tilde{S}_{\mathrm{LB}}(\mu,\lambda) = \sum_{k=0}^{L_1-1} \tilde{s}_{\mathrm{LB}}(k+\lambda S_2) w_{L_1}(k) \cdot \mathrm{e}^{-2\pi j \frac{k\mu}{L_1}} \qquad (7)$$

for $\mu \in \{0,\ldots,L_1-1\}$. This way, $S_1/S_2 = 1/\rho$ times as many analysis results are available per time unit. These can be used to produce a time-stretched (factor ρ) or pitch-scaled (factor $1/\rho$) version of the contained high band signal.

3.2. Composition of the High Band Spectrum

The high band information (DFT magnitudes for 4 – 6.4 kHz) within the upper 600 Hz of $\tilde{S}_{\mathrm{LB}}(\mu,\lambda)$ is now extracted and a (partly) synthetic DFT spectrum with L_2 bins is formed. Again, the frame index λ and the (implicit) complex conjugate symmetric extension for $\mu > \frac{L_2}{2}$ are disregarded. With $g_{\mathrm{d}} = 1/g_{\mathrm{e}}$ and μ_0, μ_1 from (4), we therefore have:

$$\left|\tilde{S}_{\mathrm{HB}}(\mu)\right| = \begin{cases} g_{\mathrm{d}} \cdot \left|\tilde{S}_{\mathrm{LB}}(\mu+\mu_0)\right| & \text{for } 0 \le \mu \le \mu_1-\mu_0 \\ 0 & \text{for } \mu_1-\mu_0 < \mu \le \frac{L_2}{2}. \end{cases} \qquad (8)$$

Compared to the DFT magnitudes, a correct representation of the phase is much less important for high-quality reproduction of higher speech frequencies, cf. [4]. In fact, there are several alternatives to obtain a suitable phase $\angle\, \tilde{S}_{\mathrm{HB}}(\mu)$. For example, an *additional* analysis of $\tilde{s}_{\mathrm{LB}}(k)$ with a window length of L_2 and a window shift of S_2 would facilitate the direct reuse of the *narrowband* phase, an approach which is often used in artificial bandwidth extension algorithms, e.g., [4]. Of course, also the *original* phase of the (pitch-scaled) high band signal could be used, if the insertion equation (4) was appropriately modified. However, the required phase post-processing (phase vocoder, see [13]) turns out to be tedious for pitch scaling by a factor of $\frac{1}{4}$ followed by a factor of 4. In fact, for our application, a simple random phase $\phi(\mu) \sim \mathrm{Unif}(-\pi,\pi)$ already delivered a high speech quality, i.e.:

$$\angle\, \tilde{S}_{\mathrm{HB}}(\mu) = \begin{cases} \angle\, \mathrm{Re}\{\tilde{S}_{\mathrm{HB}}(\mu_0)\} & \text{for } \mu = 0 \\ 0 & \text{for } \mu = \frac{L_2}{2} \\ \phi(\mu) & \text{else.} \end{cases} \qquad (9)$$

3.3. Speech Synthesis

The (partly) synthetic DFT spectrum $\tilde{S}_{\mathrm{HB}}(\mu,\lambda)$ is transformed into the time domain via an IDFT with the short window length L_2:

$$\tilde{s}_{\mathrm{HB}}(k,\lambda) = \frac{1}{L_2} \sum_{\mu=0}^{L_2-1} \tilde{S}_{\mathrm{HB}}(\mu,\lambda) \cdot \mathrm{e}^{2\pi j \frac{k\mu}{L_2}} \qquad (10)$$

for $k \in \{0,\ldots,L_2-1\}$ and 0 outside the frame interval. Now, for overlap-add, the small window shift S_2 is applied, i.e.:

$$\tilde{s}_{\mathrm{HB}}(k) = \sum_{\lambda} \tilde{s}_{\mathrm{HB}}(k-\lambda S_2,\lambda)\, w_{L_2}(k-\lambda S_2) \qquad (11)$$

for all k. With $\tilde{s}_{\mathrm{HB}}(k)$ and the corresponding low band signal $\tilde{s}_{\mathrm{LB}}(k)$, the final subband synthesis can be carried out, giving the bandwidth extended output signal $\tilde{s}_{\mathrm{BWE}}(k')$. Note that the cutoff frequency of the lowpass filter is 3.4 kHz instead of 4 kHz so that the modified components within the narrowband signal are filtered out. The remaining spectral gap between 3.4 and 4 kHz only has a very small perceptual effect as found by [14] and [4]. Example spectrograms of $\tilde{s}_{\mathrm{BWE}}(k')$ and, for comparison, $s(k')$ are shown in right part of Figure 2.

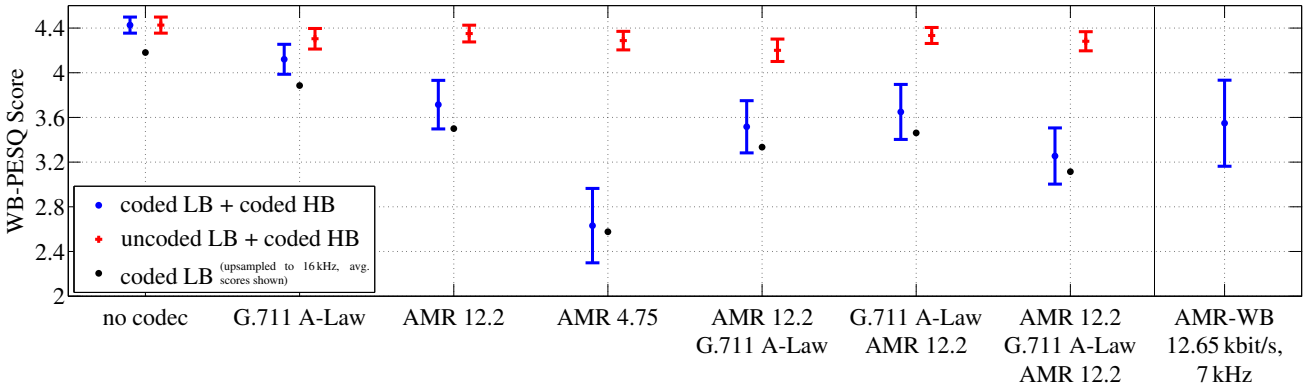

Fig. 3. Wideband speech quality (avg. WB-PESQ scores $\pm$ std. dev.) after transmission over various codecs and codec tandems.

4. QUALITY EVALUATION

Two aspects need to be considered for the quality evaluation of the proposed system. First, the narrowband speech quality must not be degraded for "legacy" receiving terminals. Second, a good (and stable) wideband quality must be guaranteed by "new" terminals according to Section 3.

Despite certain limitations w.r.t. the evaluation of bandwidth extension algorithms, the ITU-T PESQ tool [15, 16] has been used for the present evaluation. The test set comprised all American and British English speech samples of the NTT database [17], i.e., ≈ 25 min of speech.

4.1. Narrowband Speech Quality

A "legacy" terminal simply plays out the (received) composite narrowband signal $\tilde{s}_{\mathrm{LB}}(k)$. The requirement here is that the quality must not be degraded compared to conventionally encoded narrowband speech. Here, no codec has been used, i.e., $\tilde{s}_{\mathrm{LB}}(k) = s_{\mathrm{LB}}^{\mathrm{mod}}(k)$. This signal scored an average PESQ value of 4.33 with a standard deviation of 0.07 compared to the narrowband reference signal $s_{\mathrm{LB}}(k)$ which is only marginally less than the maximum achievable narrowband PESQ score of 4.55. Subjectively, it can be argued that the inserted (pitch-scaled) high frequency band induces a slightly brighter sound character that can even *improve* the perceived narrowband speech quality. This observation, however, should be substantiated with a dedicated listening test which was not conducted in the scope of the present work.

4.2. Wideband Speech Quality

A receiving terminal which is aware of the pitch-scaled high frequency content within the 3.4 – 4 kHz band can produce the output signal $\tilde{s}_{\mathrm{BWE}}(k')$ with audio frequencies up to 6.4 kHz. For a fair comparison, the reference signal $s(k')$ is lowpass filtered with the same cut-off frequency.

The wideband PESQ evaluation shows that, if no codec is used ($\tilde{s}_{\mathrm{LB}}(k) = s_{\mathrm{LB}}^{\mathrm{mod}}(k)$), an excellent score of 4.43 is obtained with a standard deviation of 0.07. Also the *subjective* listening impression confirms the high-quality wideband reproduction without any severe artifacts. However, the question remains, in how far typical codecs impair the pitch-scaled 3.4 – 4 kHz band within $s_{\mathrm{LB}}^{\mathrm{mod}}(k)$. Therefore, the ITU-T G.711 A-Law compander [18] and the 3GPP AMR codec [19, 20] at bit rates of 12.2 and 4.75 kbit/s have been chosen. Also, several codec tandems (multiple reencoding) are investigated. The respective test results are shown in Figure 3. The blue/dot markers represent the quality of $\tilde{s}_{\mathrm{BWE}}(k')$ which is often as good as (or even better than) that of AMR-WB [1, 2] at a bit rate of 12.65 kbit/s and always better than that of the corresponding narrowband signal (black markers). In contrast, the red/plus markers represent the quality that is obtained when the *original* low band signal $s_{\mathrm{LB}}(k)$ is combined with the resynthesized high band signal $\tilde{s}_{\mathrm{HB}}(k)$ *after transmission* over the codec or codec chain. This way, the quality impact on the high band signal can be assessed separately. The respective average wideband PESQ scores do not fall below 4.2 which still indicates a very high quality level.

Another short test revealed that our system is also robust against sample delays between encoder and decoder. A transmission over analog lines has not yet been tested. However, to further enhance robustness of the high band transmission, the "injection gain" g_{e} in (4) can, if necessary, still be increased without exceedingly compromising the narrowband quality.

5. DISCUSSION

The proposed system facilitates fully backwards compatible transmission of higher speech frequencies over various speech codecs and codec tandems. Its computational complexity is expected to be very moderate. The only remaining prerequisite concerning the transmission chain is that no filtering such as IRS [21] must be applied. Also, an (in-band) signaling mechanism for wideband operation is required. For instance, to maintain full backwards compatibility, a repeatedly embedded and robust 1-bit watermark can be used.

The excellent speech quality is achieved despite the heavy pitch-scaling operations because there are no dominant tonal components in the considered frequency range. Hence, a simple "noise-only" model with sufficient temporal resolution ($S_1/f_{\mathrm{s}} = 4$ ms) can be employed. Note that, if bandwidth extension towards the more common 7 kHz is desired, a pitch-scaling factor of 5 instead of 4 is unnecessary because the 6.4 – 7 kHz band can also be regenerated by fully receiver-based ABWE as, e.g., included in the AMR-WB codec [1, 2].

6. REFERENCES

[1] "ETSI TS 126 190: Adaptive multi-rate - wideband (AMR-WB) speech codec; Transcoding functions," 2001.

[2] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[3] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Edinburgh, Scotland, Sep. 1994, pp. 1178–1181.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[5] H. Pulakka, U. Rentes, K. Palomaki, M. Kurimo, and 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, Czech Republic, 2011, pp. 5100–5103.

[6] R. Taori, R. J. Sluijter, and A. J. Gerrits, "Hi-BIN: An alternative approach to wideband speech coding," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. 1157–1160.

[7] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, and S. Ragot, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, 2007.

[8] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *Proceedings of INTERSPEECH*, Lisbon, Portugal, sep 2005, pp. 1497–1500.

[9] A. Sagi and D. Malah, "Bandwidth extension of telephone speech aided by data embedding," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, January 2007, article 64921.

[10] N. Chétry and M. Davies, "Embedding side information into a speech codec residual," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep. 2006.

[11] B. Geiser and 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, Hawai'i, USA, Apr. 2007, pp. 533–536.

[12] ——, "High rate data hiding in ACELP speech codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 4005–4008.

[13] U. Zölzer, Ed., *DAFX: Digital Audio Effects*, 2nd ed. Chichester, UK: John Wiley & Sons Ltd., 2011.

[14] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1124–1137, 2008.

[15] ITU-T, "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," 2001.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[17] NTT, "NTT advanced technology corporation: Multilingual speech database for telephonometry," online: `http://www.ntt-at.com/products_e/speech/`, 1994.

[18] "ITU-T Rec. G.711: Pulse code modulation (PCM) of voice frequencies," 1972.

[19] "ETSI EN 301 704: Adaptive multi-rate (AMR) speech transcoding (GSM 06.90)," 2000.

[20] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The adaptive multi-rate speech coder," in *Proceedings of IEEE Workshop on Speech Coding (SCW)*, Porvoo, Finland, 1999, pp. 117–119.

[21] ITU-T, "ITU-T Rec. P.48: Specification for an intermediate reference system," 1976.