Steganographic Wideband Telephony Using Narrowband Speech Codecs

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

***Abstract*— We consider the transmission of wideband speech with a cut-off frequency of $f_c = 7\,\text{kHz}$ over a standardized digital narrowband communication link ($f_c = 3.4\,\text{kHz}$). At the receiver, wideband speech is produced by artificial bandwidth extension (BWE). The BWE algorithms can be realized with or without some low bit rate side information. In this paper, we propose to communicate the side information to the receiver via a steganographic channel within the bitstream of the narrowband codec. Hence, the bitstream format is not altered and the bit rate is not increased. The following codecs are considered: μ-law PCM, ADPCM, CS-ACELP, GSM FR, and GSM EFR.**

I. Introduction

The transmission of wideband speech with a cut-off frequency f_c of at least 7 kHz is a highly desirable feature for future speech/audio communication networks. Compared with conventional narrowband telephony ($f_c = 3.4\,\text{kHz}$), wideband speech offers a significantly increased subjective speech *quality* and *intelligibility* as well as a clearly reduced "*listening effort*". For wideband transmission, suitable dedicated speech codecs, such as the ITU-T G.722 or the 3GPP AMR-WB, have been developed in the past. However, the required modifications of networks and protocols turned out to be a major obstacle for the introduction of wideband speech coding in today's communication networks.

A promising approach to resolve this dilemma is the deployment of *speech bandwidth extension* (BWE), a method that (artificially) extends the limited frequency range of narrowband speech at the receiving end. The related techniques might, as anticipated in [1], be able to speed up the narrow- to wideband change-over of communication networks. In the first part of this paper (Sec. II), the *state-of-the-art* in speech bandwidth extension is reviewed briefly. We give examples for BWE algorithms that work *without* as well as *with* a certain amount of side information. BWE *with* side information is closely related to parametric speech coding and is actually an integral component of several codec standards beginning with the very first narrowband GSM Full Rate Codec [2] [3] and continuing with more recent wideband codecs such as the 3GPP Adaptive Multi-Rate Wideband Codec [4] [5] or, more explicitly, the ITU-T Embedded Variable Bit Rate Codec G.729.1 [6] [7].

A much more challenging task in speech BWE is to achieve concise results *without* transmitting any side information (see, e.g., [8]). This approach requires only modifications at the receiving end. The respective algorithms are based on the estimation of parameters of a source model for speech production given the knowledge of the narrowband signal. Unfortunately, their performance is bounded because of an insufficient amount of mutual information between the low and the high frequency subbands (cf. [9]). Yet, a certain, consistent quality improvement is achievable.

In this paper, we propose an attractive compromise between wideband speech coding with integrated BWE and purely receiver-based BWE without side information. We show how to improve BWE with a small amount of side information that is embedded into the bitstream of a narrowband codec by *steganographic techniques*. Hence, the second part of the paper (Sec. III) focuses on steganographic methods for digital speech transmission.

The third part (Sec. IV) combines speech steganography with a suitable BWE algorithm to form a transmission system that is *backwards compatible* w.r.t. legacy narrowband terminals and the network itself. The codec's bitstream format is not altered. In particular, the bit rate is not increased. The modified bitstream can be decoded by a standard narrowband decoder, possibly with a slight quality loss.

II. Speech Bandwidth Extension

Methods for extending the acoustic bandwidth of speech signals can be roughly categorized as "Bandwidth Extension with Side Information" and "Bandwidth Extension without Side Information". Exemplary algorithms for both cases are briefly reviewed below.

A. Bandwidth Extension without Side Information

Figure 1 depicts a signal flow chart of an exemplary bandwidth extension algorithm [10], [8]. This purely receiver based solution is a "mixture" of pattern recognition, statistical estimation, and speech synthesis. The algorithm exploits the implicit redundancy of the source-filter model of speech. It can be subdivided into two sub-tasks:

- extension of the spectral envelope by pattern recognition and conditional MMSE estimation
- extension of the narrowband excitation signal, e.g., by spectral replication of the base band excitation.

The narrowband speech is interpolated to 16 kHz and an *estimated wideband* linear prediction (LP) analysis filter is applied to produce the narrowband excitation. After excitation extension, the exactly inverse LP synthesis filter is applied. Therefore, the output signal contains the original narrowband

This invited paper has been presented at the *41st Asilomar Conference on Signals, Systems, and Computers* in Pacific Grove, CA, USA, Nov. 2007.

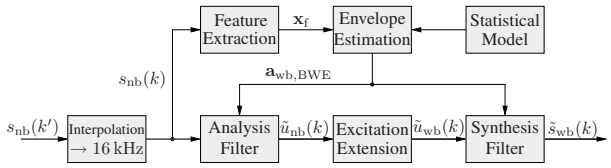

Fig. 1. Bandwidth Extension without Side Information according to [10].

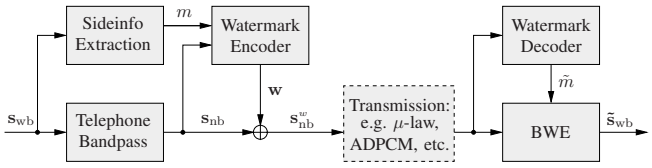

Fig. 2. BWE with signal domain data hiding (watermarking).

signal plus an artificial high band signal. Informal listening tests have confirmed that this BWE approach is consistently preferred over the narrowband speech signal.

B. Bandwidth Extension with Side Information

Owing to the reduced sensitivity characteristics of the human auditory system at higher frequencies, modern wideband codecs often synthesize these frequency components by BWE with low bit rate side information. The transmitted side information approximates the spectral and/or temporal envelopes of the high band signal which is reproduced at the decoder by applying a synthetic (e.g., noise-like) excitation signal to a corresponding time-variant filter. Recent examples are the AMR-WB [4] [5], AMR-WB+ [11], aacPlus [12], ITU-T G.729.1 [6] [7], and 3GPP2 EVRC-WB [13] [14] codecs, which realize variations and certain improvements of this basic approach.

In this paper, we use a simplified version of the TDBWE (Time Domain BandWidth Extension) module [15] which we contributed to the ITU-T G.729.1 standard. There, the amount of side information is 1.65 kbit/s. The obtained speech quality is comparable to that of true wideband coding. The simplified algorithm is described in Sec. IV-A.

III. Speech Transmission and Steganography

In principle there are three possibilities to establish a steganographic data channel within a speech signal or, respectively, its coded representation:

- *Signal Domain Data Hiding (*before *encoding)* — The respective transmission system is depicted in Fig. 2. The extracted side information is embedded into the PCM speech samples via "conventional" watermarking methods. Such transmission systems have been proposed in, e.g., [16]–[19]. A major drawback of this approach is that the steganographic transmission is not robust enough for transcoding by state-of-the-art CELP codecs.
- *Bitstream Data Hiding (*after *encoding)* — This approach modifies the bitstream, e.g., by overwriting the least significant bits, cf. Fig. 3. In contrast to the first method, the hidden data is not disturbed by the encoding process.
- *Joint coding and data hiding (*inside *the encoder)* — It is assumed that the parameter/sample is available with a resolution that is better than the resolution of the quantizer. For the transmission of N hidden bits, the set of quantizer reproduction levels is divided into 2^N subsets (binning). The transmitted information is represented by the choice of the subset. The case of scalar quantization and hidden transmission of a single bit ($N = 1$) is illustrated in Fig. 3 (requantization). The quantizer reproduction levels are divided into $2^1 = 2$ subsets with even or odd indices. As with bitstream data hiding (resulting in 101), the hidden bit is the LSB of the selected index (011), but in this case the resulting "embedding distortion" is much lower.

In the following, we will focus on the latter case and apply the respective principles to *code excited linear prediction* (CELP) speech codecs, which is today's predominant technique for low to medium-rate speech coding. Hence, we extend the principle of joint coding and data hiding to analysis-by-synthesis codecs based on vector quantization. *Joint CELP coding and data hiding* has already been investigated in [20] and a rather low steganographic capacity of 37 bit/s was achieved. A solution for state-of-the-art ACELP codecs has been introduced in [21] and it was shown that bit rates of several 100 bit/s can be reliably transmitted without compromising the quality of the coded speech signal. The respective techniques are briefly reviewed in this section.

The data hiding procedure is integrated into the fixed codebook (FCB) search of analysis-by-synthesis loop. According to the principle described above, this can be achieved by partitioning the FCB into 2^N disjoint sub-codebooks. The hidden message is *decoded* by identifying the sub-codebook that contains the received FCB vector. Considering the described embedding scheme, one might argue that the number of examined FCB entries is decreased by a factor of 2^N. The inevitable consequence would be a decreased quality of the coded speech. Yet, in practical ACELP codecs, the respective FCB search is—for reasons of complexity reduction—by far *non-exhaustive*, i.e., typically only a small, heuristically selected subset is examined during FCB search. For this reason, it is possible to establish 2^N disjoint sub-codebooks that comprise FCB entries, that have not been taken into account in the original search procedure. For properly

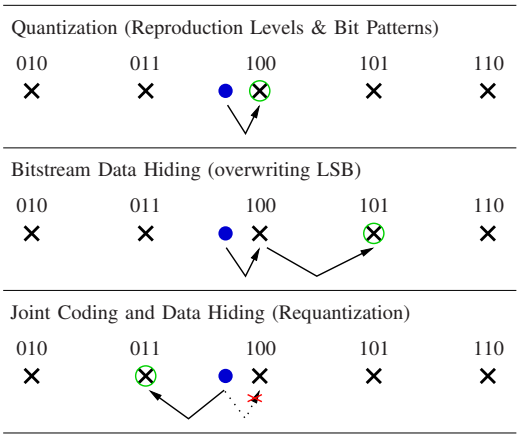

Fig. 3. Bitstream data hiding vs. joint coding and data hiding for the case of a scalar quantizer and the transmission of a side bit 1.

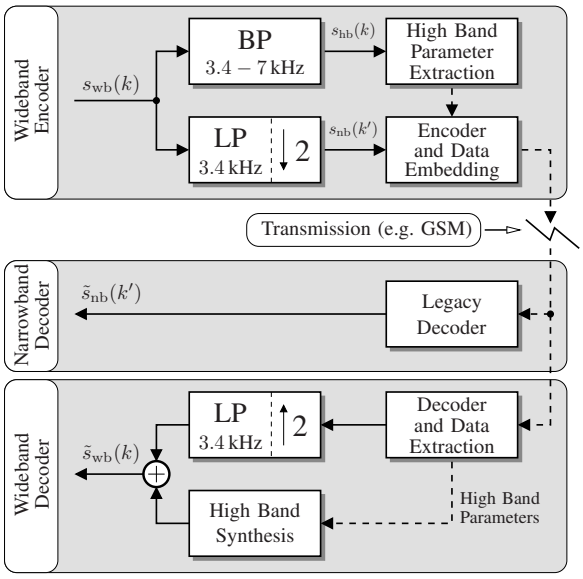

Fig. 4. Transmission system for steganographic wideband telephony.

designed sub-codebooks, the data hiding procedure does not (or only insignificantly) degrade the resulting speech quality as compared with the standard implementation.

IV. Experimental Setup

The transmission system we have used for our experiments is illustrated in Fig. 4. The input signal, which is sampled at 16 kHz, is split into a low band (LB) signal (0 – 3.4 kHz) and a high band (HB) signal (3.4 – 7 kHz). The LB signal is decimated by a factor of 2 and encoded by a narrowband speech codec. From the HB signal, a coarse parametric description (side information) is extracted for BWE at the receiver. A description of this BWE module is provided in Sec. IV-A.

The side information is communicated to the receiver using the principle of "joint coding and data hiding" as outlined in Sec. III. The data hiding has been incorporated in the following narrowband speech encoders: ITU-T G.711 [22], ITU-T G.726 [23], ITU-T G.729 [24] [25], ETSI GSM FR [2] [3], ETSI GSM EFR [26] [27]. The specific data hiding schemes for the different codecs are described in Sec. IV-B. At the receiver side, both the corresponding narrowband (legacy) decoders as well as enhanced wideband decoders with BWE have been tested. The obtained speech quality for both possibilities has been assessed and the results are presented in Sec. V.

A. BWE with 600 bit/s of Side Information

The BWE algorithm used for our experiments is depicted in Fig. 5. It is a modified version of the TDBWE module [15] of the ITU-T G.729.1 standard [6]. The received HB parameter set comprises a spectral envelope $\hat{\mathbf{F}}_{\mathrm{hb}}$ and the overall gain factor $\hat{g}_{\mathrm{hb}}$. The HB signal $\tilde{s}_{\mathrm{hb}}(k)$ is synthesized by filtering a noisy excitation $\tilde{s}_{\mathrm{exc}}(k)$ with a linear-phase time-variant FIR filter with 65 taps. The filter is constructed from the received subband gains $\hat{\mathbf{F}}_{\mathrm{hb}}$ with smoothing in time and frequency. Afterwards, the level of the output signal $\tilde{s}_{\mathbf{F}}(k)$ is adjusted by adaptive gain control. The main modifications compared with the standardized TDBWE module are:

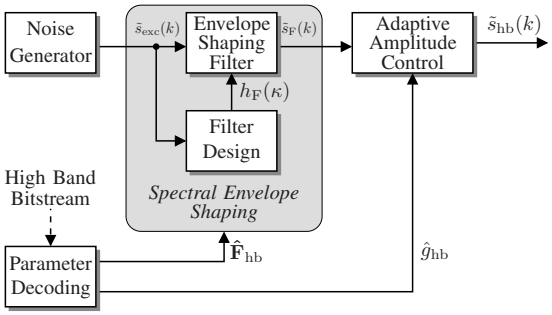

Fig. 5. Implemented algorithm for bandwidth extension with 600 bit/s of side information (simplified version of the TDBWE module [15] from the ITU-T G.729.1 codec).

- The split frequency between LB and HB is lowered from 4 kHz to 3.4 kHz.
- For simplicity, a noise excitation is used instead of a *codec specific* variable mix of harmonics and noise.
- A simplified representation of the spectral envelope of the HB signal with 5 sub-band gains is applied instead of a combination of a temporal envelope (16 sub-frame gains) and a spectral envelope with 12 sub-band gains.
- The quantization scheme has been re-designed for this new parameter set. The bit rate for side information is reduced from 1.65 kbit/s to 600 bit/s only.

For the sake of comparability of the experiments, we use this *codec independent* solution for BWE. It should be noted that there might be room for further, codec specific, improvements.

B. Data Hiding Schemes

The considered speech codecs require specific solutions for establishing the steganographic channel with a rate of 600 bit/s. The proposed solutions fulfill the following requirements:

- negligible or small degradation of the speech quality of the narrowband codecs,
- low additional computational complexity,
- and low (or zero) additional algorithmic delay.

In general, these solutions follow the principle of "joint coding and data hiding" as described above.

1) ITU-T G.711: The G.711 coder [22] is a non-uniform 8-bit scalar quantizer with 64 kbit/s. The employed compression characteristic is either A-law or μ-law. For data hiding, the current frame (40 samples) is subdivided into 3 blocks (13/13/14). One side information bit b is represented as the parity of all LSBs within a block. If this parity is not equal to b, *one* sample is requantized according to the method of "joint coding and data hiding". In particular, the sample which yields the lowest embedding distortion is selected.

2) ITU-T G.726 (ADPCM): This ADPCM codec [23] supports bit rates of 16, 24, 32, and 40 kbit/s. The normalized residual signal is logarithmically quantized within a prediction loop. "Joint coding and data hiding" is applied inside the prediction loop to 3 residual samples within a 40 sample frame. For simplicity, this is done at fixed positions. As an alternative, the parity check mechanism, which we used for the G.711 coder, could deliver somewhat better results, but requires a more complex "delayed decision" scheme.

3) ITU-T G.729 (CS-ACELP): The CS-ACELP codec [24] [25] (8 kbit/s) uses a ternary codebook structure (+1/0/-1) as CELP-FCB. Each 5 ms subframe is subdivided into 4 "tracks" (interleaved sub-grids). To form a codevector, one non-zero pulse is placed in each track. This yields 8192 possibilities for pulse positioning, but the standard only considers a small fraction. Data hiding is performed by decomposition of the *full* search space into 2^3 unique subsets.

4) ETSI GSM Full Rate (FR): Per 5 ms subframe, the Full Rate RELP codec [2] [3] (13 kbit/s) quantizes one sub-grid (RPE sequence of length 13) of a normalized residual signal using a scalar 3-bit quantizer. Three side information bits are embedded into each RPE sequence according to the parity check mechanism that we have applied for G.711.

5) ETSI GSM Enhanced Full Rate (EFR): The GSM EFR coder [26] [27] (12.2 kbit/s) uses a similar CELP-FCB structure as G.729. Therefore, the data hiding is also implemented as a decomposition of the full FCB search space into 2^3 unique subsets [21].

V. Quality Assessment

For the assessment of the resulting speech quality, we used the PESQ measure (MOS-LQO) according to ITU-T Rec. P.862.2 [28] in narrow- and wideband mode, respectively. All configurations have been tested with the complete NTT-AT database [29] (approx. 5 h of speech).

A. Narrowband Speech Quality

In order to assess the impact on the narrowband speech quality of legacy receiving terminals, the performance of the standardized narrowband decoders was compared for bitstreams with and without hidden data. The results are summarized in Fig. 6-(a). For each codec and both cases, the average MOS values ($\pm$ standard deviations) are shown side by side. The average quality loss due to data hiding varies between 0.02 (EFR) and 0.27 MOS (G.726 32 kbit/s). In informal listening tests, the quality degradations have been found to be completely (or almost) imperceptible. The somewhat lower performance for G.726 at 32 kbit/s can be explained by the low-complexity data hiding solution (cf. Sec. IV-B.2).

B. Wideband Speech Quality

The corresponding results for the "enhanced" wideband decoders are compiled in Fig. 6-(b) in three groups:

i) Our proposals for "steganographic wideband codecs" (ITU-T G.711+, ..., GSM EFR+) using BWE with side information,
ii) three reference wideband codecs (AMR-WB at 8.85 kbit/s and 12.65 kbit/s [4] [5] as well as ITU-T G.722 at 64 kbit/s [30]), and
iii) a clean LB signal with synthesized HB.

It turns out that G.729+ can deliver a quality that is comparable to AMR-WB at 8.85 kbit/s ($\Delta\mathrm{MOS} = -0.04/+0.15$), G.711+ can compete with G.722 at 64 kbit/s ($\Delta\mathrm{MOS} = -0.23/+0.12$), and GSM EFR+ yields similar results as AMR-WB at 12.65 kbit/s ($\Delta\mathrm{MOS} = -0.19/-0.04$). Thereby, the MOS standard deviation for the steganographic codecs is comparatively small. For all steganographic codecs, the input level to the narrowband core encoder was set to -22 dBov. The narrowband post-processing within the GSM EFR+ codec has been deactivated. In informal subjective tests, the listeners consistently preferred the steganographic wideband codecs over the narrowband pendants. Moreover, the codecs that use our BWE module produce clear wideband speech without noticeable artifacts.

VI. Discussion and Conclusions

In this paper, we propose a backwards compatible solution for wideband telephony that is based on widely deployed narrowband speech codecs (μ-law PCM, ADPCM, RELP, CELP) and on BWE with 600 bit/s of side information. This side information is hidden in the standard bitstream by using the principle of "joint coding and data hiding", i.e., the bitstream format is not altered and the bit rate is not increased.

The evaluation with the objective PESQ measure and informal listening tests show that the quality degradation of the narrowband speech (produced by the legacy decoders) due to data hiding is very small or negligible. The quality of the steganographic wideband codecs is rated objectively and subjectively almost as good as the reference wideband codecs at similar data rates. Potential application scenarios, which do not require any changes within the network but only upgraded terminals, include

- a compatible wideband solution for 2G and 3G cellular networks if "Tandem Free Operation" (TFO) is provided,
- a compatible wideband channel for DECT cordless telephones, and
- compatible wideband ISDN.

Our proposal could be further improved by tailoring the BWE scheme (incl. the parameter quantization) to the specific narrowband codec. Moreover, recent investigations show that, for some narrowband CELP codecs, even higher data rates (e.g, 2 kbit/s) are feasible without significant impact on the speech quality. This would give room for more precise side information, additional error protection, and signaling. Finally, it should be mentioned that the steganographic side channel could also be used for different purposes such as transmitting information that is beneficial for frame erasure concealment in packet switched networks.

Acknowledgments

The authors would like to thank *Annette Tatzel* for software development and *Tobias Breddermann* for valuable contributions.

References

[1] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?" *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, May 2006.
[2] ETSI Rec. GSM 06.10, "GSM full rate speech transcoding," version 3.2.0, Feb. 1992.
[3] P. Vary, K. Hellwig, R. Hofmann, R. J. Sluyter, C. Galand, and M. Rosso, "Speech codec for the European mobile radio system," in *Proc. of ICASSP*, vol. 1, New York, NY, USA, Apr. 1988, pp. 227–230.
[4] 3GPP TS 26.190, "AMR wideband speech codec; transcoding functions," Dec. 2001.
[5] B. Bessette, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, H. Mikkola, and K. Järvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech and Audio Proc.*, vol. 10, no. 8, pp. 620–636, Nov. 2002.
[6] ITU-T Rec. G.729.1, "G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.

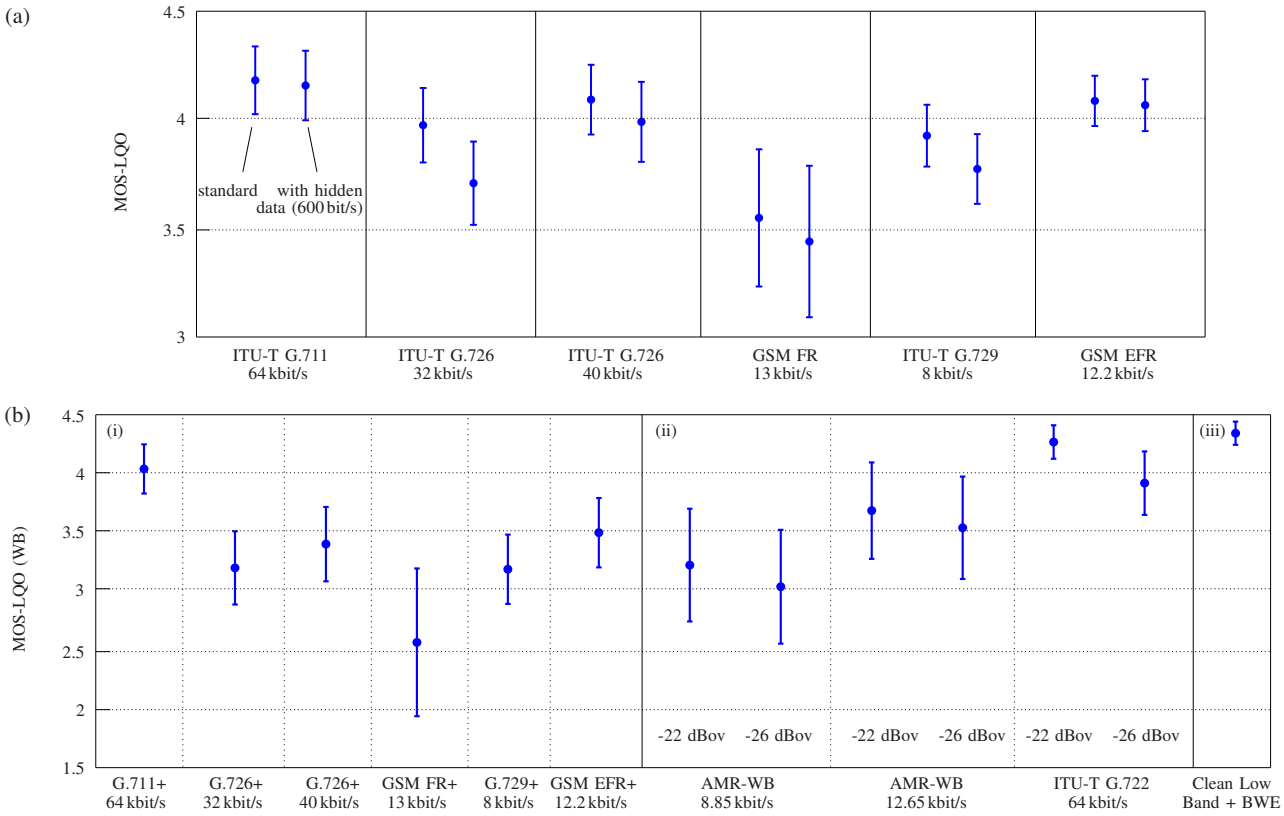

Fig. 6. Speech quality assessment results for (a) narrowband and (b) wideband decoders.

[7] S. Ragot, B. Kövesi, R. Trilling, D. Virette, N. Duc, D. Massaloux, S. Proust, B. Geiser, M. Gartner, S. Schandl, H. Taddei, Y. Gao, E. Shlomot, H. Ehara, K. Yoshida, T. Vaillancourt, R. Salami, M. S. Lee, and D. Y. Kim, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," in *Proc. of ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[8] P. Jax, "Bandwidth extension for speech," in *Audio Bandwidth Extension*, E. Larsen and R. M. Aarts, Eds. New York: John Wiley and Sons, Nov. 2004, ch. 6, pp. 171–236.

[9] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. of ICASSP*, vol. 1, Orlando, FL, USA, May 2002, pp. 237–240.

[10] ——, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[11] 3GPP TS 26.290, "Extended AMR wideband codec; transcoding functions," Sept. 2004.

[12] 3GPP TS 26.401, "Enhanced aacPlus general audio codec; general description; V7.0.0," June 2006.

[13] 3GPP2 CS0014-C v1.0, "Enhanced variable rate codec, speech service options 3, 68, and 70 for wideband spread spectrum digital systems," Feb. 2007.

[14] V. Krishnan, V. Rajendran, A. Kandhadai, and S. Manjunath, "EVRC-Wideband: The new 3GPP2 wideband vocoder standard," in *Proc. of ICASSP*, vol. 2, Honolulu, HI, USA, Apr. 2007.

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

[16] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark transmitted side information," in *Proc. of European Conf. on Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, Sept. 2005, pp. 1497–1500.

[17] H. Ding, "Wideband audio over narrowband low-resolution media," in *Proc. of ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 489–492.

[18] S. Chen and H. Leung, "Artificial bandwidth extension of telephony speech by data hiding," in *Proc. of Intl. Symp. on Circuits and Systems (ISCAS)*, Kobe, Japan, May 2005.

[19] A. Sagi and D. Malah, "Bandwidth extension of telephone speech aided by data embedding," in *EURASIP Journal on Advances in Signal Processing*, 2007.

[20] Z.-M. Lu, B. Yan, and S.-H. Sun, "Watermarking combined with CELP speech coding for authentication," *IEICE Trans. on Inf. and Systems*, vol. E88-D, no. 2, pp. 330–334, 2005.

[21] B. Geiser and P. Vary, "Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension," in *Proc. of ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[22] ITU-T Rec. G.711, "Pulse code modulation (PCM) of voice frequencies," Nov. 1988.

[23] ITU-T Rec. G.726, "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)," Dec. 1990.

[24] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," 1996.

[25] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Trans. Speech and Audio Proc.*, vol. 6, no. 2, pp. 116–130, Mar. 1998.

[26] ETSI Rec. GSM 06.60, "Digital Cellular Telecommunication System (Phase 2+); Enhanced Full Rate (EFR) speech transcoding," version 8.0.1, release 1999, Nov. 2000.

[27] K. Järvinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Adoul, "GSM enhanced full rate speech codec," in *Proc. of ICASSP*, vol. 2, Munich, Germany, Apr. 1997, pp. 771–774.

[28] ITU-T Rec. P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Nov. 2005.

[29] NTT Advanced Technology Corporation, "Multi-lingual speech database for telephonometry 1994," on-line at `http://www.ntt-at.com/products_e/speech/`.

[30] ITU-T Rec. G.722, "7 khz audio coding within 64 kbit/s," in Blue Book, vol. Fascicle III.4 (General Aspects of Digital Transmission Systems; Terminal Equipments), 1988.