PATHS TOWARD HD-VOICE COMMUNICATION

Bernd Geiser

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

ABSTRACT

In this contribution, current developments in packet-based HD-voice communication are summarized. Moreover, the, usually problematic, issue of interoperability with the already installed infrastructure is addressed. Therefore, several algorithmic approaches—including embedded coding, receiver- or network-based parameter estimation, and steganographic parameter transmission—are discussed based on the practically relevant example of parametric bandwidth extension for speech and audio signals.

Index Terms— speech coding, speech transmission, wideband, super-wideband, bandwidth extension

1. INTRODUCTION

These days, the telecommunication world is undergoing a major technology change toward a universal, packet-based network architecture for both fixed and mobile communications. The main motivations and incentives behind this effort are presumably improved flexibility and cost-efficiency. But in particular for speech and audio communication applications, the opportunity should be seized to promote high quality services which are far superior to the long-accustomed narrowband speech telephony experience. Indeed, new audio codecs, delivering additional functionality and a much better audio quality, are deployed much quicker within such a (future) network environment.

But, as a matter of fact, very little is done to improve the audio quality for today's communication networks. Instead, "least common denominator solutions" are pursued, keeping up the status quo of narrowband speech. At first sight, this might appear reasonable from the economic and marketing perspectives. However, it is nevertheless true that subscribers of new services will still experience inferior quality if their communication partner uses an old telephone or circuit-switched network access, e.g., via GSM/UMTS speech channels or private/government subnetworks. Large parts of the worldwide telephone network are in fact based on such legacy technology and can be expected to prevail for a long time. To this end, new, more advanced methods and algorithms for "High Definition" audio transmission and reproduction are required that maintain interoperability with legacy network components.

1.1. Audio Bandwidth in Voice Communication

Traditionally, telephony is conducted with a limited audio bandwidth of 0.3 – 3.4 kHz (Narrowband, NB). Two intermediate steps toward a "full band" (FB, 0.02 – 20 kHz) audio transmission are currently considered, i.e., the transmission of "wideband" speech (WB, 0.05 – 7 kHz) and, secondly, "super-wideband" (SWB) speech with frequencies up to 14 kHz. Compared to WB, SWB speech offers additional sound clarity and a "sensation of presence."

In contrast, the definition of "high definition voice" is less clear and numerous interpretations can be encountered. In the scope of this paper, "HD-Voice Communication" is understood as the transmission with an audio bandwidth of *at least* 7 kHz.

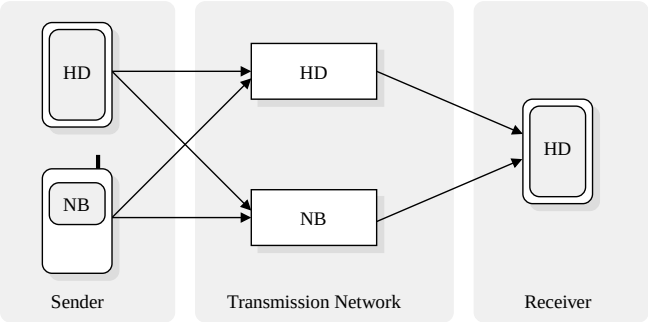


Fig. 1. Transmission scenarios in a heterogeneous network.
HD: Device with HD Audio capability
NB: Device without HD Audio capability

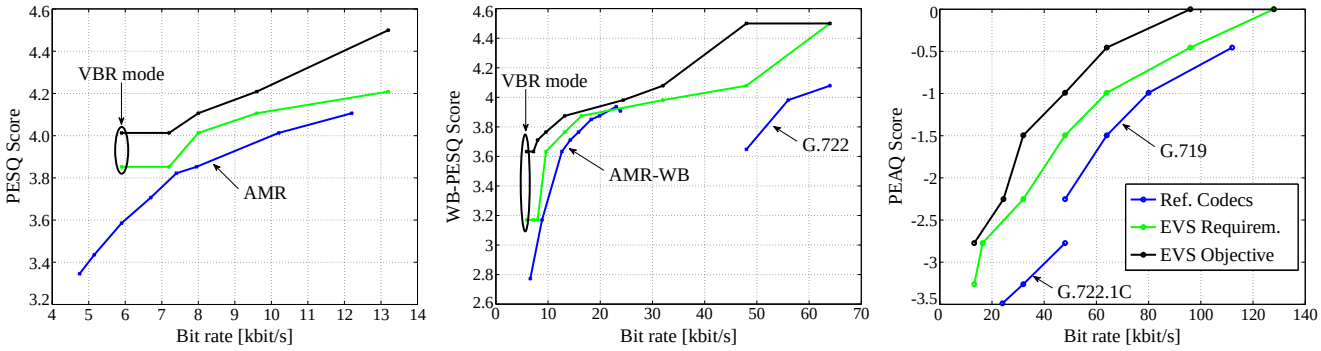
1.2. HD-Voice in a Heterogeneous Network

The "conventional" approach to establish HD-Voice communication is the deployment of suitable "HD-capable" codecs (Section 2). However, in a heterogeneous network scenario the calling terminal or a network component may still use traditional narrowband transmission, see Figure 1. Here, a dedicated HD-codec can not be applied. To ensure an HD-Voice reproduction at the receiving terminal with all conceivable transmission paths in Figure 1, various flavors of "parametric bandwidth extension" techniques can be applied, i.e., "embedded coding," "receiver/network based parameter estimation," and "steganographic parameter transmission." These approaches are discussed in Section 3. An evaluation and a comparison is provided in Section 4. The paper is summarized in Section 5.

2. DEDICATED HD-VOICE CODECS

Actually, the list of codecs that offer a transmitted audio bandwidth of 7 kHz and more and at the same time guarantee a sufficiently low algorithmic delay has become quite long. However, some candidates can be identified which can be expected to gain practical relevance, concretely:

This contribution is an accompanying paper to the keynote talk of the same title, held on September 4th at the 2012 *International Workshop on Signal Enhancement* (IWAENC) in Aachen, Germany.



(a) Narrowband (b) Wideband (c) Super-Wideband

Fig. 2. Quality Requirements for 3GPP EVS (clean speech, error free channel).

- The **ITU-T G.722** wideband codec [1] is based on a subband ADPCM algorithm. It has already been standardized in 1984 and, in the meantime, it can be used without any royalties. It is common in modern cordless telephones according to the CAT-iq standard.

- The **3GPP AMR-WB** codec [2] was standardized in 2001. It is based on the ACELP coding principle with an additional bandwidth extension from 6.4 to 7 kHz. The 12.65 kbit/s mode of this codec is increasingly applied in 3rd generation mobile networks. Also 4th generation mobile telephony (Voice over LTE, VoLTE) is foreseen to support this codec.

- The **Opus** codec was recently approved by IETF [3]. The codec, based on hybrid time and transform domain coding, is royalty-free and offers a wide range of operating conditions, i.e., audio bandwidths from narrowband to fullband over a wide range of bit rates. Additionally, in certain coding modes, the algorithmic delay can be very small. Opus is already in use in various open-source software projects (e.g., Mumble, SFLphone, etc.), but also the Skype VoIP client is planned to support Opus in the future. Moreover, the WebRTC framework implements Opus. The quality is in fact competitive with other codecs as documented by several listening tests, e.g., [4]. As Opus will be primarily used in packet-switched transmission scenarios, packet loss concealment (PLC) is also implemented.

- A possible candidate for future HD-voice communication is the 3GPP codec for *enhanced voice services* (**EVS**). The official requirements for this future codec have recently been set [5]. Similar to Opus, all relevant audio bandwidths from narrowband to (at least) super-wideband will be supported with an algorithmic delay of less than 32 ms; the computational complexity shall be less than twice the complexity of AMR-WB. The finalization of the EVS codec is targeted for the end of 2013. It is planned to include EVS in Release 12 of the 3GPP standards. As the EVS codec shall mostly be used in packet switched environment, also packet loss concealment (PLC) and, particularly, the jitter buffer management (JBM) will be integral parts of the standard. The current quality requirements for EVS are illustrated in Figure 2 for clean speech conditions based on a PESQ evaluation of the reference conditions. Basically, EVS is required to surpass all previous 3GPP and ITU-T codecs in terms of quality.

3. PARAMETRIC BANDWIDTH EXTENSION

In the following, three different approaches for HD-Voice communication in a *heterogeneous network* will be summarized which are based on the concept of *parametric bandwidth extension* (BWE) as illustrated in Figure 3. It is assumed that a band-limited speech signal is available at the decoder side. Then, additional frequency content (mostly toward higher frequencies) is regenerated based on a compact parametric description. This approach is justified because the human auditory perception is rather insensitive to spectral details at higher audio frequencies. In fact, only certain "coarse" signal characteristics must be preserved which can effectively be described by parametric coding techniques. The approaches to be summarized in the following obtain their concrete parameter sets in different ways and are applicable in different transmission scenarios in the heterogeneous network model of Figure 1.

3.1. Embedded Coding

In embedded coding [6], a bitstream "extension layer" is appended to the bitstream of the existing (narrowband) codec. For parametric BWE, typical add-on bit rates vary between 0.5 and 5 kbit/s, depending on the signal type (speech or general audio) and on the split frequency, i.e., NB to WB or WB to SWB. Embedded coding is only applicable in the HD→HD→HD scenario according to Figure 1 but, in contrast to a dedicated (and incompatible) HD codec, codec renegotiation and transcoding can be avoided here as the extension layer can simply be discarded at any time. Meanwhile, numerous standardized codecs are available that explicitly implement BWE techniques, e.g., AMR-WB+, EVRC-WB, HE-AAC, MPEG USAC, G.719 as well as G.718/G.729.1 and their SWB annexes. The BWE parameter sets typically comprise temporal and spectral envelopes and, in some cases, a compact description of certain spectral details such as harmonic and tonal signal components.

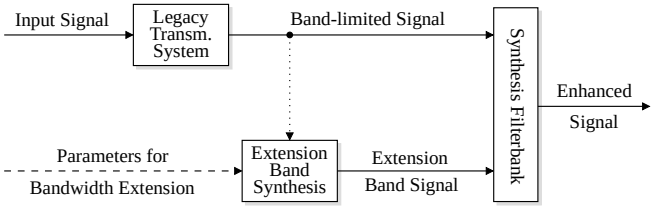


Fig. 3. System for parametric bandwidth extension (BWE) of band-limited speech or audio signals.

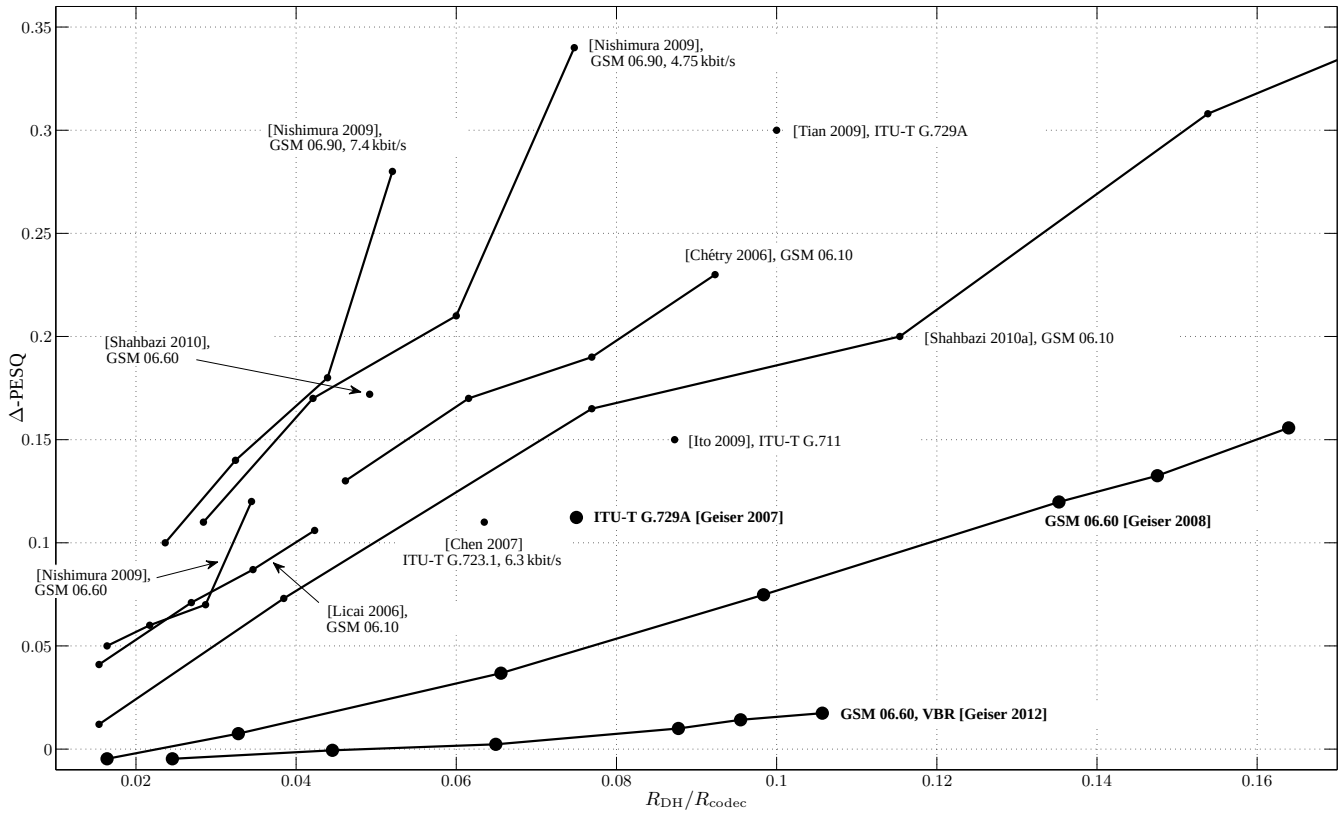


Fig. 4. Comparison of speech codec data hiding algorithms based on a relative hidden bit rate ($R_{\mathrm{DH}}/R_{\mathrm{codec}}$). Evaluation of average PESQ loss compared to standard codec (Δ-PESQ). PESQ values are taken from the respective publications.

3.2. Receiver-Based Parameter Estimation

If a narrowband *sending* terminal is used in Figure 1, a BWE parameter set is not readily available and statistical estimation techniques need to be applied instead, either in the receiver (NB→NB→HD), or in the network (NB→HD→HD). The respective algorithms, also known as *Artificial BWE* (ABWE), estimate the required parameter set based on certain *features* from the baseband signal with the help of a pre-trained statistical model. However, such models only available for known source characteristics, i.e., for speech signals. As another interesting application, an estimated ABWE signal can be inserted if the extension layer of an embedded codec was discarded temporarily. This is, e.g., applied in the G.729.1-SWB codec.

For ABWE systems, consistent quality advantages over NB speech transmission have been reported, e.g., [7]. Recently, using ABWE, also an improved *speech intelligibility* could be found, at least in noisy environments [8]. Also, first investigations of ABWE to extend WB speech to the SWB bandwidth reveal promising results as shown in Section 4.

3.3. Steganographic Parameter Transmission

If, in Figure 1, only the terminals are replaced to support HD-Voice reproduction, but the network is not adapted (HD→NB→HD), a *hidden* transmission of the BWE parameter set can be considered where the related information is hidden in the bitstream of the NB codec using steganographic techniques. A suitably equipped receiving terminal can exploit the hidden information to regenerate the missing audio frequencies. Most importantly, the full compatibility with old (NB) receiving terminals is maintained. For example, for the widely

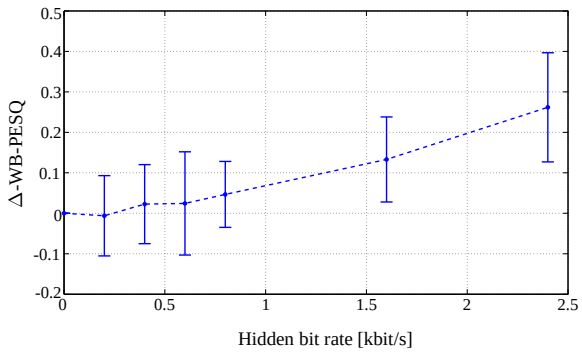
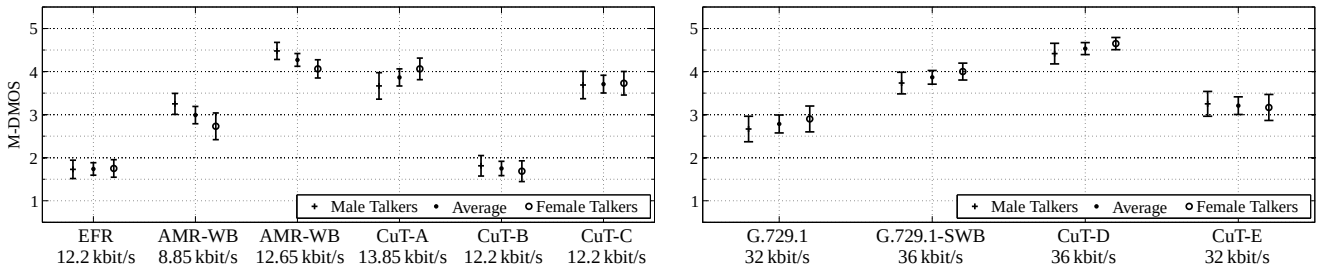


Fig. 5. Data hiding in the AMR-WB codec.

used ACELP speech codecs, the data hiding scheme of [9] can be used which offers a sufficient hidden bit rate to support the transmission of BWE parameters (e.g., 2 kbit/s within the GSM EFR codec).

In the described application, it is important that the hidden data will not compromise the NB quality, i.e., the "digital watermark" must not be audible. This is analyzed in Figure 4 for various speech data hiding algorithms based on an average PESQ-score loss (Δ-PESQ). The VBR algorithm labeled with [Geiser 2012] is identical to the original algorithm of [9] (labeled [Geiser 2008]) except that certain, quality-sensitive speech frames have been identified and a lower hidden bit rate has been used therein. Figure 5 analyzes the application of this algorithm to the AMR-WB codec thus facilitating a bandwidth extension toward the SWB bandwidth. It must be noted that even a Δ-PESQ-value of 0.15 has been found to be subjectively barely noticeable in a listening experiment, see [9].



(a) Wideband (b) Super-Wideband

Fig. 6. Results of the super-wideband DCR test (95% conf. intervals).

4. EVALUATION & COMPARISON

The concepts described above have been assessed in two subjective listening tests, i.e., for NB to WB extension as well as for WB to SWB extension. The test samples were compared with the original (uncoded) version based on a *modified* DMOS scale: degradation is inaudible (5), barely audible (4), clearly audible but not annoying (3), slightly annoying (2), annoying (1). In total, 96 votes were received per test condition.

4.1. Wideband Speech Quality

For the WB case, test results are shown in Figure 6(a). Apart from the NB anchor (EFR) and the WB references (AMR-WB), the following codecs have been assessed:

- **CuT-A: Embedded Coding** — A BWE parameter set is determined at the encoder and *quantized* with 1.65 kbit/s. This information is appended to the 12.2 kbit/s bitstream of the 3GPP EFR codec, thus forming an *embedded codec* with a sum bit rate of 13.85 kbit/s. The quality approaches that of AMR-WB.
- **CuT-B: Receiver Based Parameter Estimation** — Bayesian estimation with Hidden-Markov modeling is used to estimate the BWE parameters based on features of the NB signal. CuT-B does not increase the bit rate. In the present, rather strict M-DMOS test, CuT-B cannot improve quality over EFR. However, in a direct AB comparison of both codecs, CuT-B is clearly preferred.
- **CuT-C: Steganographic Parameter Transmission** — ACELP data hiding with a hidden bit rate of 1.65 kbit/s is employed to transport the quantized BWE parameter set in a backwards compatible manner. The bit rate of CuT-C is still 12.2 kbit/s and the quality is almost identical to that of CuT-A. Only for female voices, a slight degradation can be observed.

4.2. Super-Wideband Speech Quality

For the SWB case, test results are shown in Figure 6(b). The ITU-T G.729.1 codec is used as WB anchor. The SWB annex of this codec is used as SWB reference. The following codecs have been tested:

- **CuT-D: Embedded coding** — The BWE parameters are *quantized* with a bit rate of 4 kbit/s. This information is appended to the bitstream of the G.729.1 codec, thus forming an *embedded codec* with 36 kbit/s. CuT-D is identical with the 36 kbit/s mode of "candidate B" for G.729.1-SWB standardization, see [10]. The quality is clearly better than that of the SWB reference.
- **CuT-E: Parameter estimation** — Again, Bayesian estimation with Hidden-Markov modeling is used to estimate the SWB parameter set based on features of the WB signal. CuT-E does not increase the bit rate. Compared to the WB anchor, an improved quality could be shown here.

The steganographic AMR-WB codec (Figure 5) has not been tested here.

5. SUMMARY

Several approaches for HD-voice reproduction/transmission have been discussed based on a heterogeneous network scenario Figure 1. The achieved speech quality has been assessed in subjective listening tests both for wideband and super-wideband speech signals. More details on the tested algorithms as well as on the concrete test conditions and results can be found in [11].

6. REFERENCES

[1] ITU-T Recommendation 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.

[2] 3GPP TS 26.190, "AMR wideband speech codec; transcoding functions," Dec. 2001.

[3] IETF RFC 6716, "Definition of the Opus Audio Codec," Sept. 2012.

[4] A. Rämö and H. Toukomaa, "Voice Quality Characterization of IETF Opus Codec," in *Proc. of INTERSPEECH*, Florence, Italy, 2011.

[5] 3GPP, "EVS Permanent Documents," 2012, online: http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/.

[6] B. Geiser, S. Ragot, and H. Taddei, "Embedded Speech Coding: From G.711 to G.729.1," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds., chapter 8, pp. 201–247. John Wiley and Sons, Ltd., Chichester, UK, Jan. 2008.

[7] 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 *Proc. of IEEE ICASSP*, Prague , Czech Republic, 2011, pp. 5100–5103.

[8] F. Heese, B. Geiser, and P. Vary, "Intelligibility assessment of a system for artifical bandwidth extension of telephone speech," in *Proc. of DAGA*, Mar. 2012, pp. 905–906.

[9] B. Geiser and P. Vary, "High rate data hiding in ACELP speech codecs," in *Proc. of IEEE ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4005–4008.

[10] B. Geiser et al., "Candidate proposal for ITU-T super-wideband speech and audio coding," in *Proc. of IEEE ICASSP*, Tapei, Taiwan, Apr. 2009, pp. 4121–4124.

[11] B. Geiser, *High-Definition Telephony over Heterogeneous Networks*, Dissertation, Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN), Vol. 33, Verlag Mainz in Aachen, June 2012, also available online: http://darwin.bth.rwth-aachen.de/opus3/volltexte/2012/4184/pdf/4184.pdf.