22. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), September 2011, Aachen, Germany (TuDPress, ISBN 978-3-942710-37-4)

Binaural Wideband Telephony Using Steganography

Bernd Geiser, Magnus Schäfer, and Peter Vary

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

Abstract: A system for the transmission of binaural wideband speech signals over a standard telephone network is proposed. It is backwards compatible with the (single channel and narrowband) 3GPP Adaptive Multirate (AMR) codec. The required information about the *source location* and for *audio bandwidth extension* is transmitted over a steganographic communication channel that is embedded within the bitstream of the AMR codec. A legacy receiver can still decode the single channel narrowband signal without noticeable quality loss.

1 Introduction

The reproduction of binaural wideband speech signals in speech communication systems allows a much more natural user experience than traditional telephony. Therefore, high-quality conversational speech codecs are needed that not only provide a higher acoustical bandwidth than the traditional system, but also reproduce binaural (2-channel) signals. This demand has, for example, been addressed within 3GPP by standardizing the AMR-WB+ codec [12, 1]. However, the introduction of a new codec into an established communication systems often breaks backwards compatibility. An interoperable solution is obtained by adding enhancement layers to a standardized codec. Moreover, it is also possible to *hide* these enhancement bits in the bitstream by using *data hiding* techniques, leading to a full *backwards compatibility*.

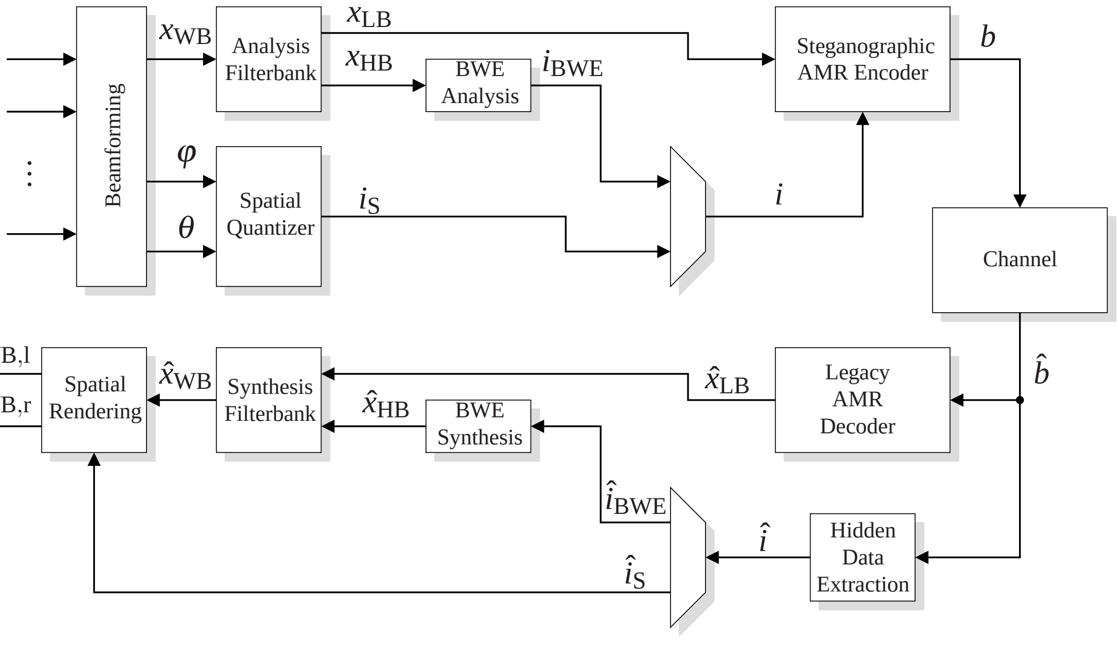

Figure 1 - System for binaural wideband telephony using a steganographic AMR codec.

In this paper, a data hiding technique for the 3GPP AMR codec [6, 4] (as deployed in today's GSM and UMTS cellular networks) is used to transmit enhancement layers that allow to *widen the transmitted bandwidth* and that facilitate a *spatial rendering* at the decoder side. Hence, not only a higher audio quality is achieved, but also the ability to localize the sound source is provided in a backwards compatible manner. A *legacy* terminal (with its standard decoder) will ignore the hidden information and reproduce the standard (narrowband) speech output without noticeable degradations (compared to a narrowband reference).

1.1 System Overview

A block diagram of the proposed transmission system is depicted in Figure 1. It is based on the ACELP (algebraic CELP) data hiding mechanism from [9] which allows to hide steganographic data with 2 kbit/s in the bitstream of the 12.2 kbit/s mode of the AMR codec. At the encoder side, source location information i_{S} is generated by a multi-microphone beamformer. The wideband input speech is split into two frequency bands. The lower band signal x_{LB} is encoded by the AMR codec and parameters i_{BWE} for bandwidth extension are extracted from the higher band signal x_{HB}. Both parameters are multiplexed and transmitted over the steganographic channel, i.e., the bit rate of the AMR codec is not increased. The decoder performs bandwidth extension and spatial rendering based on the received information.

1.2 Organization of the Paper

In the following, first, the ACELP data hiding mechanism is detailed (Section 2). Then, the employed bandwidth extension (Section 3) as well as spatial acquisition and rendering techniques (Section 4) are specified. The paper concludes with an example application scenario (Section 5).

2 AMR Speech Coding with Hidden Data

This section reviews the ACELP (algebraic CELP) data hiding mechanism from [9] which allows to hide steganographic data with 2 kbit/s $= 40$ bit/frame in the bitstream of the 12.2 kbit/s mode of the AMR codec [6, 4].

In order to maintain the speech quality of the coder, the steganographic bits are embedded in less important parts of the AMR bitstream, i.e., in the fixed codebook (FCB) contribution of the codec. The impact of the hidden bits on the speech quality is minimized by a *joint* implementation of the speech encoding and data hiding operations, cf. [8, 14]. The key to this "ACELP steganography" is a modified search strategy for the ACELP codebook.

First, we first need to define the "message" that shall be embedded into a 5 ms subframe. The index i in Figure 1, which is determined for every 20 ms frame of the input signal, is split such that a particular steganographic message m corresponds to 10 individual bits of i. Each message m (to be hidden in the respective 5 ms subframe) is therefore given as a 10 bit binary sequence which is, again, split into five sub-messages with two bits each. The sub-messages are denoted by, e.g. $(m)_{0,1}$ for the first two bits of m.

To enable the transmission of $N = 10$ steganographic bits, the ACELP codebook (or fixed codebook, FCB) is partitioned into $M = 2^{10}$ *sub-codebooks* that uniquely identify the selected message m. Based on the standard ACELP search method from [6], the proposed steganographic algorithm has been derived in two steps: Codebook Partitioning and Search Space Expansion.

2.1 Codebook Partitioning

The M disjoint sub-codebooks are established by appropriately restricting the set of admissible codevectors. In particular, a specific parity condition is imposed on certain parts of the AMR bitstream:

$$(m)_{2k,2k+1} = \left[\mathscr{G}\left(\left\lfloor \frac{i_k}{5} \right\rfloor\right) \oplus \mathscr{G}\left(\left\lfloor \frac{i_{k+5}}{5} \right\rfloor\right)\right] \bmod 4, \qquad (1)$$

for the ACELP pulse positions i_k with $k \in \{0,\ldots,4\}$. $X \oplus Y$ is the bitwise exclusive disjunction (XOR) of two binary strings and $\mathscr{G}$ represents the standardized Gray encoding of the ACELP pulse position codewords. At the decoder, the hidden information can be retrieved directly from the AMR bitstream using Equation (1).

2.2 Search Space Expansion

Based on the chosen codebook partitioning, an FCB search strategy can be devised that provides a good trade-off between speech quality and computational complexity. Thereby, the “admissible values” for the pulse positions i_{k+5} can be computed by solving Equation (1) for i_{k+5}. The limitation in admissible pulse positions is compensated by an extended search space. Concretely, *quadruples* of pulse positions are optimized instead of position *pairs* as in the standard codebook search algorithm. More details on this steganographic FCB search can be found in [9].

3 Bandwidth Extension

The TDBWE algorithm, which is used here to perform bandwidth extension of the narrowband AMR signal (0.05 – 4 kHz) towards the wideband frequency range (0.05 – 7 kHz), is standardized as a part of ITU-T Rec. G.729.1 [10, 13]. However, it is also easily applied to the 3GPP AMR codec, see for instance [11].

At the encoder side, a fairly coarse parametric description of the high frequency components (4 – 7 kHz) of the 20 ms input signal frames is computed. The respective parameter set comprises temporal and spectral energy envelopes, concretely:

- A time envelope consisting of 16 subframe gains $T(i)$. The subframe length is 1.25 ms. This resolution is chosen to concisely represent sounds like plosives in speech signals.

- A frequency envelope consisting of 12 subband energies $F(i)$. The frequency envelope is computed for every 20 ms frame. It is interpolated at the decoder to reduce the number of parameters to be quantized and to get a smooth envelope every 10 ms. The physical frequency bandwidth is 375 Hz.

The 28 TDBWE parameters are quantized with a bit rate of 1.65 kbit/s. The employed method is mean-removed split VQ. The mean time envelope M_T is also transmitted. The time envelope is quantized in 2 equal blocks with 8 parameters each, while the frequency envelope is quantized in 3 equal blocks with 4 parameters each. The vector codebooks are trained using a modified K-means algorithm forcing centroids on a rectangular grid. The concrete bit allocation for the TDBWE bitstream is detailed in the upper part of Table 1. The TDBWE bits are converted into the index i_{BWE} as shown in Figure 1.

At the decoder side, first, a so called “excitation signal” is synthetically generated based on information from the narrowband layers of the respective baseband codec (ITU-T G.729 or 3GPP

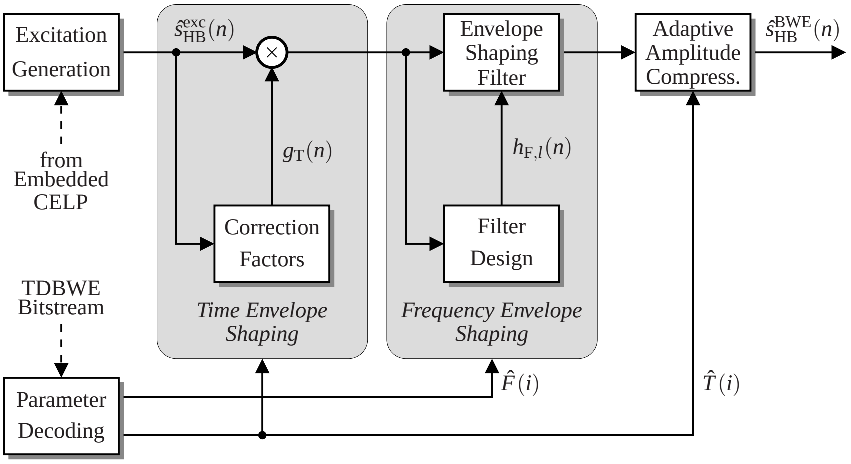

Figure 2 - Decoder of the TDBWE bandwidth extension algorithm from ITU-T Rec. G.729.1.

AMR). The excitation signal is a weighted mixture of noise and periodic components. The latter are produced by an overlap-add of spectrally shaped and suitably spaced glottal pulses. Then, its time and frequency envelopes are consecutively shaped by gain manipulations and filtering operations to match the transmitted parametric description. Contrary to classical LPC-based BWE methods, the TDBWE model reconstructs the higher band by shaping an artificial excitation signal according to a desired time envelope (energy per time segments) and a desired frequency envelope (energy per subbands). Time envelope shaping is implemented as a sample-based multiplication by a gain factor, while frequency shaping is performed using a bank of linear-phase finite impulse response (FIR) filters with 2 ms delay. Finally, a post-processing procedure attenuates residual artifacts. The TDBWE decoder is shown in Figure 2. A comprehensive and complete description of TDBWE is provided in [7] and in the text of the ITU-T G.729.1 recommendation [10].

4 Spatial Acquisition and Rendering

The transmission of spatial information that is used here relies on a separation of the source signal x_{WB} itself from information about the direction. The source direction is represented here by the azimuth angle φ and the elevation angle θ which is mapped to the closest available source position present in the chosen set of binaural impulse responses $h_{j,L|R}$.

In the encoder, both angles are jointly quantized by a spatial quantizer resulting in the quantization index i_{S} and embedded into the AMR bitstream as described in Section 2. At the decoder, this index is retrieved from the hidden bitstream by the hidden data extraction and a separation from the received quantization index $\hat{i}_{\mathrm{BWE}}$ of the bandwidth extension part as described in Section 3. The received spatial index $\hat{i}_{\mathrm{S}}$ is then utilized to address the predefined set of binaural impulse responses and the impulse responses $h_{\hat{i}_{\mathrm{S}},L}$ and $h_{\hat{i}_{\mathrm{S}},R}$ are selected for the binaural synthesis.

The binaural synthesis is done in the time domain and consists of a frame-wise filtering of the reconstructed wideband signal $\hat{x}_{\mathrm{WB}}$ with the binaural impulse responses. To avoid filter switching artifacts, a short crossfade between the filter coefficients in successive frames is used.

Table 1 - Example bit allocation for the steganographic bitstream (40 bit per 20 ms = 2 kbit/s).

Parameter	Symbol	Dimension	# bits
mean time envelope	M_T	1	5
mean-removed time envelope (1)	$\mathbf{T}_\mathbf{1}^\mathrm{M}$	8	7
mean-removed time envelope (2)	$\mathbf{T}_\mathbf{2}^\mathrm{M}$	8	7
mean-removed frequency envelope (1)	$\mathbf{F}_\mathbf{1}^\mathrm{M}$	4	5
mean-removed frequency envelope (2)	$\mathbf{F}_\mathbf{2}^\mathrm{M}$	4	5
mean-removed frequency envelope (3)	$\mathbf{F}_\mathbf{3}^\mathrm{M}$	4	4
Azimuth	φ	1	7
Elevation	θ	1	0
Sum	Σ	31	40

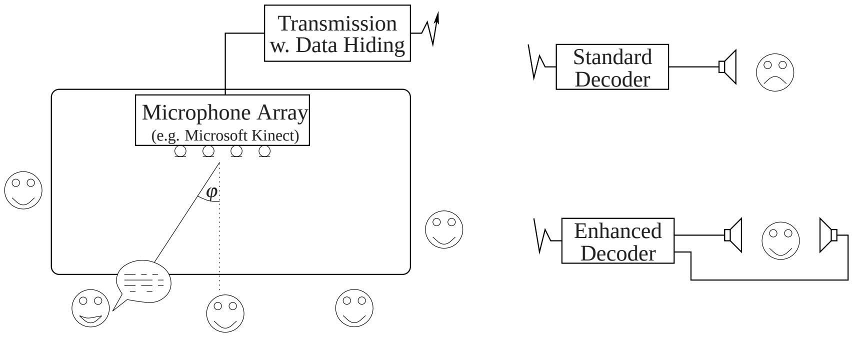

Figure 3 - Example application scenario: Conference with two external participants (downlink only).

5 Example Application Scenario

A typical application scenario for the proposed transmission system is illustrated in Figure 3. In this conference setting, a microphone array isolates the active speaker. The respective speech signal and the detected angle φ are supplied to the transmission system (see Figure 1). An enhanced decoding unit which is aware of the hidden information can reproduce a binaural wideband signal. In contrast, a standard decoder outputs plain narrowband telephone speech. Note that the elevation θ is not used in this scenario, thus leaving room for a more accurate representation of the azimuth angle φ.

The bit allocation of the steganographic bitstream which is used in the present scenario is shown in Table 1. Apart from the 33 bits per 20 ms which are used for bandwidth extension, 7 bits are reserved to encode the angle φ. The binaural impulse responses for the spatial rendering in the present application are a subset from the continuous impulse response measurements described in [5] and [2]. The subset consists of 127 pairs of impulse responses (addressed with $i_\mathrm{S} = 0, \ldots, 126$) covering the frontal half of the horizontal plane with an angular resolution of 1 degree between -36 and 36 degrees (with 0 degrees being directly in front) and a resolution of 2 degrees between -90 and -38 as well as 38 and 90 degrees. This non-uniform resolution was chosen due to the fact that the human hearing system exhibits a higher resolution in frontal directions compared to lateral directions [3]. The remaining unused index value ($i_\mathrm{S} = 127$) can be utilized to (temporarily) switch off the binaural rendering.

6 Conclusions

The proposed system for binaural wideband communication provides a significantly enhanced user experience compared to standard mobile telephony without compromising interoperability with deployed transmission equipment. It has been shown that even low additional data rates (e.g., 2 kbit/s), if economically used, suffice to introduce multiple additional features into a speech communication system in a backwards compatible manner.

References

[1] 3GPP TS 26.290: *Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions*. 2004

[2] Antweiler, C. ; Enzner, G. : Perfect Sequence LMS for Rapid Acquisition of Continuous-Azimuth Head Related Impulse Responses. In: *Proc. of IEEE WASPAA*. New Paltz, NY, USA, Oct. 2009, pp. 281–284

[3] Blauert, J. : *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, 1997. – ISBN 9780262024136

[4] Ekudden, E. ; Hagen, R. ; Johansson, I. ; Svedberg, J. : The adaptive multi-rate speech coder. In: *Proc. of IEEE Speech Coding Workshop*. Porvoo , Finland, 1999, pp. 117–119

[5] Enzner, G. : Analysis and Optimal Control of LMS-Type Adaptive Filtering for Continuous-Azimuth Acquisition of Head Related Impulse Responses. In: *Proc. of IEEE ICASSP*. Las Vegas, NV, USA, Mar. 2008, pp. 393–396

[6] ETSI Recommendation GSM 06.90: *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding*. version 7.2.1, release 1998, Apr. 2000

[7] Geiser, B. ; Jax, P. ; Vary, P. ; Taddei, H. ; Schandl, S. ; Gartner, M. ; Guillaumé, C. ; Ragot, S. : Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1. In: *IEEE Tr. Audio, Speech, and Language Proc.* 15 (2007), Nov., No. 8, pp. 2496–2509

[8] Geiser, B. ; Vary, P. : Backwards Compatible Wideband Telephony in Mobile Networks: CELP Watermarking and Bandwidth Extension. In: *Proc. of IEEE ICASSP*. Honolulu, Hawai'i, USA, Apr. 2007

[9] Geiser, B. ; Vary, P. : High Rate Data Hiding in ACELP Speech Codecs. In: *Proc. of IEEE ICASSP*. Las Vegas, NV, USA, Mar. 2008

[10] 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

[11] Jung, S.-K. ; Ragot, S. ; Lamblin, C. ; Proust, S. : An embedded variable bit-rate coder based on GSM EFR: EFR-EV. In: *Proc. of IEEE ICASSP*, 2008, pp. 4765–4768

[12] Makinen, J. ; Bessette, B. ; Bruhn, S. ; Ojala, P. ; Salami, R. ; Taleb, A. : AMR-WB+: a new audio coding standard for 3rd generation mobile audio services. In: *Proc. of IEEE ICASSP*. Philadelphia, PA, USA, Mar. 2005

[13] Ragot, S. et al.: 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 IEEE ICASSP*. Honolulu, Hawai'i, USA, Apr. 2007

[14] Vary, P. ; Geiser, B. : Steganographic Wideband Telephony Using Narrowband Speech Codecs. In: *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*. Pacific Grove, CA, USA, Nov. 2007, pp. 1475–1479. – Invited Talk