HIERARCHICAL MULTI-CHANNEL AUDIO CODING BASED ON TIME-DOMAIN LINEAR PREDICTION

Magnus Schäfer and Peter Vary

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

ABSTRACT

A novel hierarchical multi-channel coding scheme is proposed which exhibits a significant decrease in decoding complexity compared to earlier proposals.

The new coding scheme is based on a single channel downmixing process followed by predictions of the multi-channel input signals. Symmetries in the prediction filter coefficients and the prediction errors allow for a reduced number of channels which need to be transmitted.

A detailed evaluation of the achievable prediction gain and the impact of quantization on the perceived quality leads to insights into the appropriate choice of system parameters. Besides the attractive feature of being usable as a hierarchic extension to existing single channel communication systems and its very low additional algorithmic delay, the transmission quality of the proposed design also scales very well with the available data rate.

Index Terms— Multi-Channel Coding, Linear Prediction, Hierarchical

1. INTRODUCTION

Efficient coding of stereo or multi-channel signals with low algorithmic delay in a heterogeneous system environment is a topic of growing interest due to the fact that even mobile communication devices nowadays offer multiple microphones and two (when used with a headset) or even multiple loudspeakers. Multiple microphones are so far mostly utilized for signal enhancement, e.g., noise reduction or dereverberation. The possibility of headphones to reproduce spatial information is not exploited by current communication systems.

Predictive coding systems that exploit the temporal and spatial correlations between the individual channels are advantageous for this scenario as they usually have very low algorithmic delay and low computational complexity.

The earliest proposals that can be seen as a simple compressive time-domain predictive encoding scheme for multi-channel signals (i.e., not simply by using multiple single channel systems) date back to the 1950s and 1960s when FM broadcasting was extended to allow for the transmission of stereo signals [1]. This system takes correlation between the two channels into account by means of a fixed preprocessing step which generates one sum channel to ensure backwards compatibility to mono receivers.

An overview on more recent developments in time-domain predictive coding of multi-channel signals can be found in [2]. A different approach for the coding of stereo signals is to tightly integrate the stereo prediction into the codec. One example is the complex-valued stereo prediction as proposed in [3] for the Unified Speech and Audio Coding (USAC) approach.

The developments in the area coding of multi-channel signals for scenarios like storage or streaming (e.g., MPEG Surround [4]) which are not critical with respect to algorithmic delay are summarized in [5]. The algorithms from these approaches were recently reconfigured for low-delay operation and tailored for a combination with the AAC-ELD codec as presented in [6].

One important point for any coding scheme that shall be deployed in the existing telephone as well as in a high-quality audio conferencing network should be to ensure backwards compatibility. For the single channel case, many known codecs achieve this by being structured in a hierarchical manner (e.g., [7] or [8]). In contrast to that, the aforementioned recent approaches for multi-channel coding do not incorporate any possibility for a single channel receiver when combined with codecs that are in use in the telephone network. The presented approach is usable with any mono codec for the main channel ensuring backwards compatibility (an example combination of an earlier version of the proposed coding scheme with the Adaptive Multi-Rate Wideband [9] can be found in [10]).

In the remainder of this paper, the coding system will first be introduced and a comparison of the decoding complexity between the previous and the new system will be given. An evaluation of the performance follows that sheds light on the influence of quantization on the achievable transmission quality.

2. CODING SYSTEM

The proposed multi-channel coding system is a generalization and improvement of the two-channel system proposed in

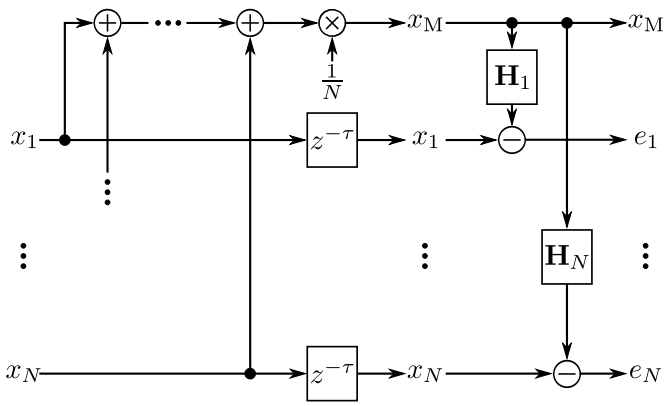

Fig. 1. Encoding Structure of the Multi-Channel Coding System

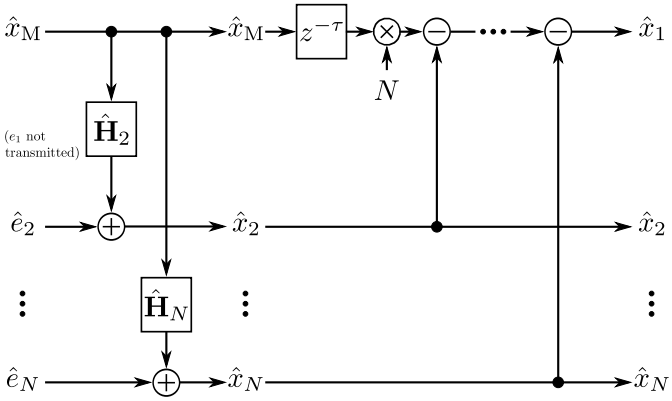

Fig. 2. Decoding Structure of the Multi-Channel Coding System

[11]. It consists of low-complexity downmixing of all N input channels $x_n(k)$ into one so-called main channel $x_M(k)$ according to

$$x_{\mathrm{M}}(k) = \frac{1}{N} \cdot \sum_{n=1}^{N} x_n(k)\,. \tag{1}$$

From this main channel, a prediction of delay compensated versions of the individual input channels is carried out as depicted in Fig. 1. In the original proposal, this prediction was carried out by linear phase finite impulse response (FIR) filters to exploit amplitude relations between the main channel and the input channels. For general multi-channel signals, we propose to remove the linear phase constraint which allows to fully utilize both amplitude and phase relations to increase the prediction gain of the encoding stage. The prediction errors e_n are then calculated as

$$e_n(k) = x_n(k-\tau) - \sum_{\lambda=0}^{L-1} h_n(\lambda) \cdot x_{\mathrm{M}}(k-\lambda). \tag{2}$$

Therein, τ is the delay of the delay elements that are present in Fig. 1 and the filter length amounts to $L = 2\cdot\tau+1$. This choice for the filter length leads to a symmetric structure: The delayed sample $x_n(k-\tau)$ in the input channel is predicted from the sample in the main channel with identical delay $x_{\mathrm{M}}(k-\tau)$ and from both τ newer and τ older samples. The filter coefficients $h_n(\lambda)$ are derived by a minimum mean square error (MMSE) criterion

$$E\left\{e_n^2(k)\right\} \rightarrow \min. \tag{3}$$

The derivation of the expected value of the squared prediction error with respect to the filter coefficients leads to a form that is very similar to the regular normal equations known from single channel linear prediction, e.g., [12]:

$$\mathbf{X}_{\mathrm{MM}} \cdot \mathbf{H}_n = \mathbf{X}_{n\mathrm{M}} \tag{4}$$

Just like in the single channel case, the matrix $\mathbf{X}_{\mathrm{MM}}$ contains the autocorrelation values $\varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(\lambda)$ of the main channel in symmetric Toeplitz structure while the column vector $\mathbf{H}_n$ is composed of the filter coefficients $h_n(\lambda)$:

$$\mathbf{X}_{\mathrm{MM}} = \begin{pmatrix} \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(0) & \cdots & \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(L-1) \\ \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(1) & \cdots & \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(L-2) \\ \vdots & \ddots & \vdots \\ \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(L-1) & \cdots & \varphi_{x_{\mathrm{M}}x_{\mathrm{M}}}(0) \end{pmatrix} \tag{5}$$

$$\mathbf{H}_n = \begin{pmatrix} h_n(0)\ h_n(1)\ \ldots\ h_n(L-1) \end{pmatrix}^{\mathrm{T}} \tag{6}$$

The difference to the single channel case lies in the column vector $\mathbf{X}_{n\mathrm{M}}$. This vector contains the cross correlation values $\varphi_{x_n x_{\mathrm{M}}}(\lambda)$ between the respective input channel n and the main channel:

$$\mathbf{X}_{n\mathrm{M}} = \begin{pmatrix} \varphi_{x_n x_{\mathrm{M}}}(\tau)\ \varphi_{x_n x_{\mathrm{M}}}(\tau-1)\ \ldots\ \varphi_{x_n x_{\mathrm{M}}}(-\tau) \end{pmatrix}^{\mathrm{T}} \tag{7}$$

When calculating the filter coefficients $\mathbf{H}_n$ according to Eq. (4) and the prediction error signals e_n for all N channels and then using Eq. (1), some advantageous symmetries can be found:

- The sum of all vectors of filter coefficients equals a vector with only zeros and a single one in the middle:

$$\sum_{n=1}^{N} \mathbf{H}_n = \begin{pmatrix} \underbrace{0\ \ldots\ 0}_{\tau} & 1 & \underbrace{0\ \ldots\ 0}_{\tau} \end{pmatrix}^{\mathrm{T}} \tag{8}$$

- The sum of all prediction errors equals zero at all times:

$$\sum_{n=1}^{N} e_n(k) = 0 \tag{9}$$

Based on these findings, the sum signal and only $N-1$ prediction errors and $N-1$ sets of filter coefficients have to be calculated and transmitted to the decoder as the missing values can be reconstructed by applying Eqs. (8) and (9). This can be done in an efficient way by a simplified decoding structure compared to [11] which is depicted in Fig. 2 for the case $N = 2$.

Therein, the reconstruction of all output channels but one is achieved by an inversion of the prediction process from the encoder:

$$\hat{x}_n(k-\tau) = \hat{e}_n(k) + \sum_{\lambda=0}^{L-1} \hat{h}_n(\lambda) \cdot \hat{x}_{\mathrm{M}}(k-\lambda). \qquad (10)$$

The signal $\hat{x}_1$ (w.l.o.g) is then calculated using

$$\hat{x}_1(k-\tau) = N \cdot \hat{x}_{\mathrm{M}}(k-\tau) - \sum_{n=2}^{N} \hat{x}_n(k-\tau) \qquad (11)$$

It can directly be seen that this structure allows for a perfect reconstruction of the input signals given that the transmission introduces no errors (i.e., $\hat{x}_{\mathrm{M}} = x_{\mathrm{M}}$, $\hat{e}_n = e_n$ and $\hat{\mathbf{H}}_n = \mathbf{H}_n$).

An overview of the decoding complexity of the new structure in comparison to the original decoder from [11] is given in Tab. 1.

	Previous decoder	New decoder
Multiplications	$N \cdot L$	$(N-1) \cdot L + 1$
Additions	$(N+1) \cdot L$	$(N-1) \cdot (L+1)$

Table 1. Complexity of the previous and new decoding structure

It can be seen that the new decoder requires fewer multiplications as long as there are any filters $\mathbf{H}_n$ in the system, i.e., $L > 0$ and it also requires fewer additions as long as the filters are reasonably long in comparison to the number of channels, i.e., $L > \frac{N-1}{2}$ which is a sensible setup as will be seen later in the evaluation.

For the transmission of stereo signals, i.e., $N = 2$ (currently probably the most important practical use-case) with a filter length of $L = 11$, the number of multiplications and summations is reduced by 45% and 64%, respectively.

3. EXPERIMENTS

Experiments were carried out using the 3GPP audio dataset [13] consisting of approximately ten minutes of clear and noisy speech from various talkers in different languages as well as music signals. All signals are sampled at $f_s = 48$ kHz. The number of channels is two for the entire dataset, hence only one set of filter coefficients and one prediction error has to be transmitted besides the main channel.

An overview on the resulting stereo coding system can be found in Fig. 3. The filter coefficients $\mathbf{H}_2$ are assumed to be transmitted transparently with negligible errors (i.e., $\hat{\mathbf{H}}_2 = \mathbf{H}_2$) while the main channel x_{M} and the prediction error e_2 are subject to quantization with the quantizers $\mathcal{Q}_{\mathrm{M}}$ and $\mathcal{Q}_2$, respectively. A logarithmic scalar quantizer according to the μ-law characteristic [12] with $\mu = 255$ is used with varying word length w from 1 to 12 bit. Additionally, the length L of the prediction filter is varied from 1 to 50 taps to evaluate the impact of these variables on the overall performance of the coding system. Note that the high prediction filter lengths are solely used for illustrative purposes here as they would lead to correspondingly high data rates when used in a practical system.

Two measures will be utilized to quantify the performance of the coding system:

- Average prediction gain between the input signals and the prediction errors (K represents the total number of samples in the signals):

$$\overline{G}_p = \frac{1}{2} \cdot \sum_{n=1}^{2} 10 \cdot \log_{10} \left(\frac{\sum_{k=1}^{K} x_n(k)^2}{\sum_{k=1}^{K} e_n(k)^2} \right) \qquad (12)$$

  This definition of the prediction gain differs slightly from the usually employed formula due to the fact that the prediction here is carried out between two signals instead of within one signal. For a practical implementation, only one prediction has to be carried out due to the symmetries described in Eqs. 8 and 9. For the evaluation, the prediction is done for both channels here.

- Average and minimum Perceptual Evaluation of Audio Quality (PEAQ) [14] values of the entire transmission chain. The two-channel mode of PEAQ is used as a measure for the subjective quality. When used with two-channel signals, PEAQ first calculates all the model output variables of its perceptual model individually for the two channels which quantify the degradation of $\hat{x}_1$ and $\hat{x}_2$ with respect to x_1 and x_2, respectively. These output variables are then averaged linearly and fed into an artificial neural network to calculate the final Objective Difference Grade (ODG) value. The PEAQ scale ranges from 0 (i.e., degradation is imperceptible) to -4 (i.e.,

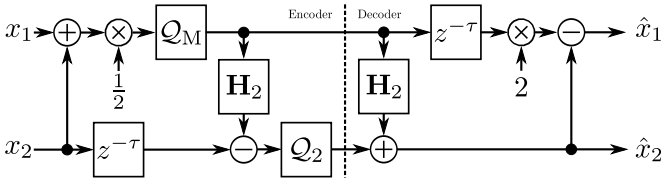

Fig. 3. Stereo encoding and decoding system used for the experiments

degradation is very annoying) and a value of -2 or better is very acceptable for most use cases.

In order to lay the focus of the evaluation on the performance of the predictive coding step, both quantizers $\mathcal{Q}_{\mathrm{M}}$ and $\mathcal{Q}_2$ are neither specifically trained nor adaptive to the signal in any way. The logarithmic scalar quantizers use the well-known μ characteristic (e.g., [12]) for companding and a uniform symmetric mid-tread quantizer as their core quantizer. This uniform quantizer utilizes $2^w - 1$ quantization levels (i.e., a word length of 1 bit means that all values are quantized to zero) and it is designed to cover the entire possible range of values of the input data x_n (i.e., -1 to 1 for this data set).

4. RESULTS

The relation between the length L of the prediction filters, the achievable prediction gain and the transmission quality is depicted in Fig. 4. It can be seen that even fairly short lengths offer significant prediction gains of more than 11 dB. This gain increases with increasing filter length but the increase in gain starts to get smaller fairly quickly, even a filter length of 50 taps increases the prediction gain only by roughly another 2 dB.

The minimum PEAQ value increases as well for an increasing length of the prediction filter indicating an increased robustness of the transmission system for longer filter lengths. The simulations for this graph were carried out for a perfect transmission of the main channel and with varying word lengths for the quantizer $\mathcal{Q}_2$ of 8 and 9 bit and their respective minimum PEAQ values were then averaged. Taking both measures into account, a filter length of about 25 taps appears to be reasonable to ensure a good performance of the coding system. However, this choice would lead to a significantly increased data rate for the transmission of the filter coefficients. As a compromise between a good performance of the system and a low necessary data rate, a value of $L = 11$ is chosen for the other evaluations. This equates to an additional algorithmic delay $\frac{\tau}{f_s}$ of about $0.23\,\mathrm{ms}$.

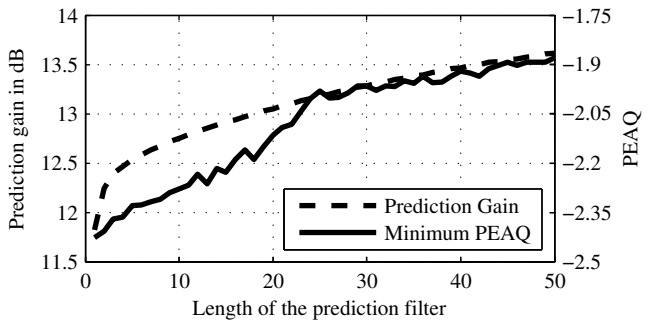

Fig. 4. Prediction Gain and PEAQ values for different filter lengths.

The achievable prediction gain for $L = 11$ is depicted in Fig. 5 over the word length of the quantizer for the main channel. The dashed line represents the prediction gain for the unquantized case as a reference.

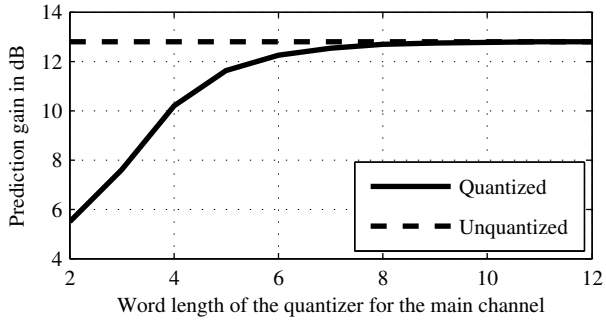

Fig. 5. Average prediction gain for different word lengths of the quantizer $\mathcal{Q}_{\mathrm{M}}$ for the main channel.

It can be seen that the performance of the prediction step depends on the chosen quantizer $\mathcal{Q}_{\mathrm{M}}$ for the main channel. Word lengths of less than 5 bit lead to a significant performance decrease in this setup compared to the prediction gain for the unquantized case.

The achievable quality for the transmission system of Fig. 3 is illustrated in Fig. 6. The abscissa therein is the word length w_2 of the quantizer for the prediction error while the set of curves consists of the different word lengths w_{M} of the quantizer for the main signal (2,3,5,7 and 10 bit from bottom to top). The impact of the word length of $\mathcal{Q}_2$ is obviously bigger than the impact of the word length of $\mathcal{Q}_{\mathrm{M}}$ which can be explained by the fact that any quantization error within $\hat{e}_2$ will be present in both $\hat{x}_1$ and $\hat{x}_2$ without any filtering.

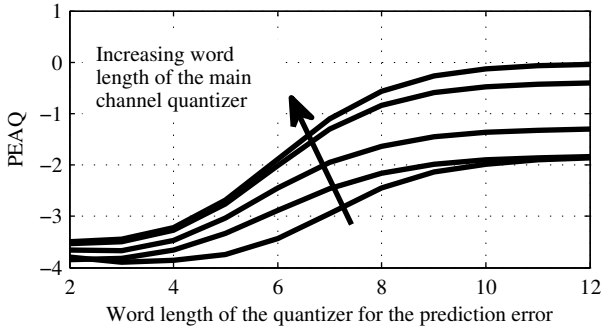

Fig. 6. Average PEAQ values for different word lengths of the quantizers for the main channel $\mathcal{Q}_{\mathrm{M}}$ and the prediction error $\mathcal{Q}_2$. The set of curves depicts word lengths w_{M} of 2, 3, 5, 7, and 10 bit from bottom to top.

As a reference, the perceptual quality of a symmetric, independent transmission of x_1 and x_2 is depicted in Fig. 7. This system uses one logarithmic quantizer for each channel

that are both set to identical word lengths. It can be seen that a longer word length leads to a higher quality and that a word length of 7 bit for each quantizer leads to a PEAQ value of approximately -2 (which is very acceptable for many use cases).

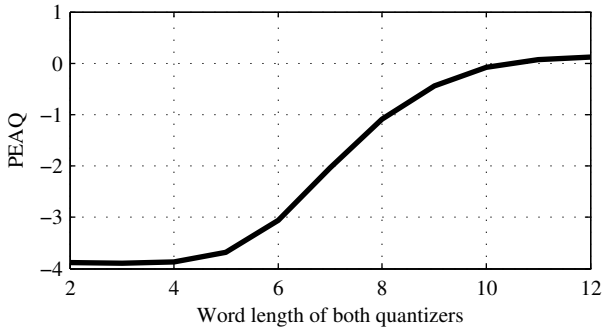

Fig. 7. Average PEAQ values for different word lengths of the quantizers when using a symmetric transmission without the prediction structure.

A comparable quality for the proposed predictive coding scheme can already be reached for a combination of $w_{\mathrm{M}} = 2$ bit and $w_2 = 10$ bit. Even including an unoptimzed and very precise quantization of the filter coefficients at 16 bit per coefficient, the overall data rate for the proposed structure amounts to just 88 percent of the rate that is necessary for the independent transmission.

The same findings can also be made when using a single channel codec, e.g., [15], instead of the logarithmic quantizer. However, the performance evaluation as presented is more transparent with respect to the understanding of the proposed audio coding structure.

5. CONCLUSIONS

A generalized version of the linear predictive coding scheme from [11] was presented which allows to apply the concept to multi-channel signals while also significantly decreasing the computational complexity.

The presented encoding structure achieves good prediction gains at reasonably short filter lengths. A detailed evaluation of the impact of quantization on the overall system performance shows that the proposed audio coding system achieves a very good transmission quality at higher data rates and that there is a graceful degradation in PEAQ performance through medium data rates.

From the PEAQ evaluation, it can also be deduced that the data rate efficiency for the proposed system is significantly higher than for an independent transmission of the input channels which uses the same quantizers. More sophisticated coding techniques can be utilized to further decrease the absolute data rate.

The additional algorithmic delay that is introduced by the system is very low which makes it an attractive possibility for the transmission of multi-channel signals in a backwards compatible way to existing mono systems.

6. REFERENCES

[1] Federal Communications Commission, "Amendment of Part 3 of the Commission's Rules and Regulations to Permit FM Broadcast Stations to Transmit Stereophonic Programs on a Multiplex Basis," 1961.

[2] Arijit Biswas, *Advances in Perceptual Stereo Audio Coding Using Linear Prediction Techniques*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[3] C.R. Helmrich, P. Carlsson, S. Disch, B. Edler, J. Hilpert, M. Neusinger, H. Purnhagen, N. Rettelbach, J. Robilliard, and L. Villemoes, "Efficient Transform Coding of Two-Channel Audio Signals by Means of Complex-Valued Stereo Prediction," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011.

[4] ISO/IEC, *Information technology – Coding of audio-visual objects – Part 3: Audio*, International Organization for Standardization, 2009.

[5] Jürgen Herre, "From joint stereo to spatial audio coding - recent progress and standardization," in *Proceedings of the 7th Int. Conference on Digital Audio Effects*, 2004.

[6] Manfred Lutzky, María Luis Valero, Markus Schnell, and Johannes Hilpert, "AAC-ELD V2 - The New State of the Art in High Quality Communication Audio Coding," in *Audio Engineering Society Convention 131*, 10 2011.

[7] 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.

[8] ITU-T Rec. G.718, "Frame Error Robust Narrowband and Wideband Embedded Variable Bit-Rate Coding of Speech and Audio from 8-32 kbit/s," 2008.

[9] ITU-T Rec. G.722.2, "Wideband Coding of Speech at Around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," 2003.

[10] Magnus Schäfer, Hauke Krüger, and Peter Vary, "Extending Monaural Speech and Audio Codecs by Inter-Channel Linear Prediction," in *20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, 2009, vol. 1.

[11] Hauke Krüger and Peter Vary, "A New Approach for Low-Delay Joint-Stereo Coding," in *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, Oct. 2008.

[12] P. Vary and R. Martin, *Digital Speech Transmission – Enhancement, Coding and Error Concealment*, John Wiley & Sons, Inc., Chichester, UK, 2006, ISBN 0-471-56018-9.

[13] 3GPP, "Speech Codec Speech Processing Functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) Speech Codec; Conformance Testing," TS 26.274, 3rd Generation Partnership Project (3GPP), June 2007.

[14] ITU, *Method for Objective Measurements of Perceived Audio Quality (ITU-R Recommendation BS.1387-1)*, International Telecommunications Union, 2001.

[15] ITU-T Rec. G.726, "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," 1990.