EXTENDING MONAURAL SPEECH AND AUDIO CODECS BY INTER-CHANNEL LINEAR PREDICTION

Magnus Schäfer, Hauke Krüger, and Peter Vary

Institut für Nachrichtengeräte und Datenverarbeitung (IND), RWTH Aachen, 52056 Aachen
E-Mail: {schaefer,krueger,vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract:

In this contribution, we propose the application of a novel concept for a flexible hierarchical stereo extension of existing monaural speech and audio codecs. The concept is based on inter-channel linear prediction of the left and right channels from a sum signal and allows for a very flexible extension of existing speech and audio codecs. In contrast to theoretical examinations in earlier publications, a stereo codec is built in this contribution by combining the new stereo framing with the core transmission of the standardized Adaptive Multi Rate - WideBand codec (AMR-WB) in a hierarchical manner. The proposed modification introduces just marginal additional system delay compared to mono AMR-WB.

It will be shown by simulations that compared to an individual transmission of left and right channel, the application of the inter-channel linear prediction concept achieves an identical quality at a significantly lower data rate for an important class of stereo signals. This is due to a concentration of most of the signal energy in the sum signal and the filter coefficients while the prediction error is of lesser importance for the quality. This will be shown by gradually decreasing the transmission data rate for the prediction error to the point of a purely parametric solution where only the sum signal and the filter coefficients are transmitted.

1 Introduction

Broadcasting of stereophonic signals started already in 1961. The basis for Frequency Modulated (FM) stereo broadcasting is the production of a mid (for compatibility with existing mono receivers) and a side channel signal (M/S stereo) from the left and right channel signals. In each modulated FM radio channel, the mid channel signal is transmitted in the baseband spectrum and the side channel signal in the spectrum related to the amplitude modulated *double-sideband suppressed carrier signal* (DSSCS, [10], [14]). Still nowadays, FM radio receivers may reconstruct either only the monaural mid channel representation (mono) of the input stereo signal from only the baseband spectrum, or the complete stereo image signal if also the DSSCS signal is demodulated. In digital audio compression, a lot of confusion is related to the term joint-stereo coding. In the literature, it is referred to as both, M/S and Intensity Stereo coding. The target of joint-stereo coding is to enable a higher compression ratio in a joint coding approach in comparison to an approach in which the signals for left and right channel are coded independently.

A lot of joint-stereo approaches in the literature are based on a high resolution frequency domain representation of the input signal (e.g. Intensity Stereo Coding, [3],[9]) and therefore subject to a high algorithmic delay. In contrast to these techniques, joint-stereo coding approaches in the time domain more easily achieve low algorithmic delay. In [6], an adaptive

inter-channel predictor is proposed that is composed of an inter-channel FIR prediction filter and a delay. Predictor filter coefficients and inter-channel delay adapt to the given signals for left and right channel. The target of this approach is to produce an estimate of the first channel on the basis of the second channel to reduce the signal variance of the predicted channel and hence save bits. Adaptive multichannel prediction is also investigated in [13] and revisited in [2]. In this case, inter- and intra-channel predictors are optimized in a joint way to produce residual signals with reduced signal variance in both channels to reduce the overall bit rate for lossless coding. Similar concepts were also investigated for packet loss concealment in multichannel environments in [12]. Both techniques are not suitable to extend existing mono codecs in a hierarchical way.

In this paper, the application of an alternative approach for joint-stereo coding is proposed. The concept operates in the time domain and enables low algorithmic delay while being usable as a hierarchical extension to any given mono codec. Compared to the approaches given in the literature, the sum of left and right stereo input signal (the mono representation) is filtered by linear phase FIR filters to predict the left *and* the right channel. Due to its modularity, the new approach is suitable to extend existing monaural codecs toward the coding of stereo signals while preserving backwards compatibility with monaural transmission.

The new approach is especially well suited for signals recorded with coincident recording techniques like X-Y or M/S microphone setups as described, e.g., in [4].

2 The Hierarchical Linear Prediction Stereo Extension

The proposed new approach has some similarities to the usual M/S joint-stereo encoding principle. Therefore, M/S joint-stereo will later be considered as a benchmark for the performance of the prediction-based extension. Like M/S stereo, the new approach operates in the time domain. The reference block diagram for the development of the new stereo transmission is shown in Figure 1. First, a mono signal is calculated from the right and the left channel input signal which

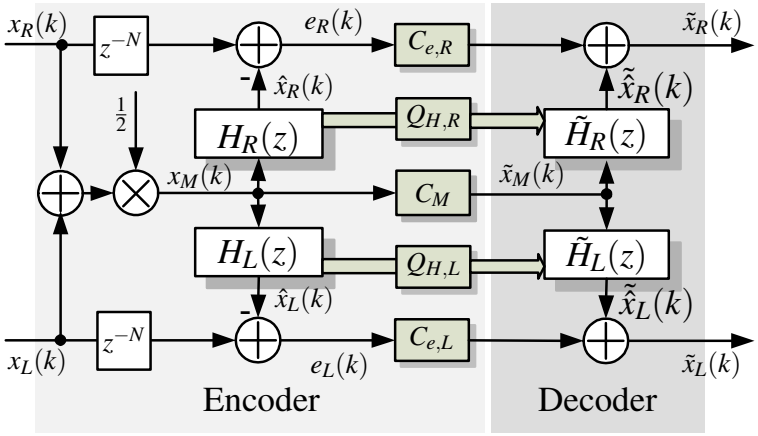

Figure 1 - Reference block diagram for the new approach for joint-stereo coding.

is identical to the mid signal as known from M/S joint-stereo coding,

$$x_M(k) = \frac{x_R(k) + x_L(k)}{2}. \tag{1}$$

However, the side channels are produced in a different manner: Where conventional M/S joint-stereo coding uses a plain difference of the input signals for its side channel, the new approach applies linear filters with system functions $H_L(z)$ and $H_R(z)$ respectively to generate estimates

$\hat{x}_L(k)$ and $\hat{x}_R(k)$ for the left and right channel input signals. These estimates are subtracted from the left and right channel input signals which results in the channel-specific prediction error signals $e_L(k)$ and $e_R(k)$. This filtering approach allows to exploit higher-order correlation as will be seen in the derivation of the optimum filter coefficients in 2.1.

The filters $H_L(z)$ and $H_R(z)$ are symmetric linear phase FIR filters with $(2 \cdot N + 1)$ filter coefficients,

$$\begin{aligned} H_L(z) &= a_L(0) \cdot z^{-N} + \sum_{i=1}^{N} a_L(i) \cdot (z^{-N-i} + z^{-N+i}) \\ H_R(z) &= a_R(0) \cdot z^{-N} + \sum_{i=1}^{N} a_R(i) \cdot (z^{-N-i} + z^{-N+i}). \end{aligned} \tag{2}$$

The stereo residual signals $e_L(k)$ and $e_R(k)$ are computed as the difference between a delayed version of the input signals $x_L(k)$ and $x_R(k)$ and the estimate signals $\hat{x}_L(k)$ and $\hat{x}_R(k)$,

$$\begin{aligned} e_L(k) &= x_L(k-N) - a_L(0) \cdot x_M(k-N) - \\ &\quad \sum_{i=1}^{N} a_L(i) \cdot (x_M(k-N-i) + x_M(k-N+i)) \\ e_R(k) &= x_R(k-N) - a_R(0) \cdot x_M(k-N) - \\ &\quad \sum_{i=1}^{N} a_R(i) \cdot (x_M(k-N-i) + x_M(k-N+i)). \end{aligned} \tag{3}$$

Delaying the input signals is required to compensate for the delay introduced by the linear phase filters. This encoding scheme leads to five different signals that need to be transmitted to the decoder. In addition to the mono signal $x_M(k)$ and the residual signals $e_L(k)$ and $e_R(k)$ that are encoded and decoded with appropiate mono codecs C_M, $C_{e,L}$ and $C_{e,R}$, two sets of $(N+1)$ stereo prediction coefficients $a_L(i)$ and $a_R(i)$ are quantized with the quantizers $Q_{H,L}$ and $Q_{H,R}$ and transmitted independently as depicted in Figure 1.

2.1 Optimal Filter Coefficients

For the calculation of the optimal stereo prediction filter coefficients $a_L(i)$ and $a_R(i)$, it is assumed that the signals $x_L(k)$ and $x_R(k)$ are stationary over one block of the speech codec to be used. Without loss of generality, we will carry out the derivation on the right channel signal and transfer the results to the left channel later.

The target of the optimization procedure is to minimize the expectation of the squared residual signal $e_R(k)$:

$$E\{e_R^2(k)\} \rightarrow \min \tag{4}$$

To improve readability, we will use the substitution

$$a'_R(i) = \begin{cases} \frac{1}{2} \cdot a_R(i) & \text{for } i = 0 \\ a_R(i) & \text{for } i > 0 \end{cases} \tag{5}$$

for the following calculations. With equation (3) and setting its partial derivatives with respect to all $a'_R(i)$ zero yields the following equation:

$$\mathbf{X}_M \cdot \mathbf{a}'_R = \mathbf{X}_{R,M}. \tag{6}$$

The vector

$$\mathbf{a}'_R = \begin{bmatrix} a'_R(0) & a'_R(1) & \cdots & a'_R(N) \end{bmatrix}^T \tag{7}$$

contains the desired filter coefficients. The matrix

$$\mathbf{X}_M = \begin{bmatrix} X_M(0,0) & \cdots & X_M(0,N) \\ \cdots & X_M(j,l) & \cdots \\ X_M(N,0) & \cdots & X_M(N,N) \end{bmatrix} \tag{8}$$

is composed of the autocorrelation function values φ_{x_M,x_M} related to the mono signal $x_M(k)$,

$$X_M(j,l) = \varphi_{x_M,x_M}(\mid l-j \mid) + \varphi_{x_M,x_M}(\mid l+j \mid) \tag{9}$$

with the index l and j to address columns and rows respectively. The vector $\mathbf{X}_{R,M}$ consists of the cross correlation function values,

$$\mathbf{X}_{R,M} = \begin{bmatrix} (\frac{\varphi_{x_R,x_M}(0)+\varphi_{x_R,x_M}(-0)}{2}) \\ (\frac{\varphi_{x_R,x_M}(1)+\varphi_{x_R,x_M}(-1)}{2}) \\ \cdots \\ (\frac{\varphi_{x_R,x_M}(N)+\varphi_{x_R,x_M}(-N)}{2}) \end{bmatrix}. \tag{10}$$

The optimal filter coefficients $\mathbf{a}'_R$ are hence

$$\mathbf{a}'_R = (\mathbf{X}_M)^{-1} \cdot \mathbf{X}_{R,M} \tag{11}$$

for the right channel signal. The filter coefficients for the left channel are determined in analogy to (6)-(11) as

$$\mathbf{a}'_L = (\mathbf{X}_M)^{-1} \cdot \mathbf{X}_{L,M}. \tag{12}$$

2.2 Simplification of the coding concept

With the equations to determine the optimal filter coefficients and the relation

$$\varphi_{x_R,x_M}(i) + \varphi_{x_L,x_M}(i) = 2 \cdot \varphi_{x_M,x_M}(i), \tag{13}$$

due to (1) it can be shown that

$$\begin{aligned} \mathbf{a}'_R + \mathbf{a}'_L &= (\mathbf{X}_M)^{-1} \cdot (\mathbf{X}_{R,M} + \mathbf{X}_{L,M}) \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T. \end{aligned} \tag{14}$$

Accordingly, there is a very simple relation between the coefficients for the left and the right channel. In analogy to this, with (3) and (14), a simple relation can be derived for the residual signals for left and right channel as well,

$$e_L(k) + e_R(k) = 0 \; \forall \; k. \tag{15}$$

Considering this result, Figure 1 can be transformed into the diagram shown in Figure 2. As a result, only the filter coefficients $\mathbf{a}_R$, the residual signal $e_R(k)$ and the mono signal $x_M(k)$ must be transmitted to reproduce the left and the right channel signal in the decoder which reduces the required overall bit rate.

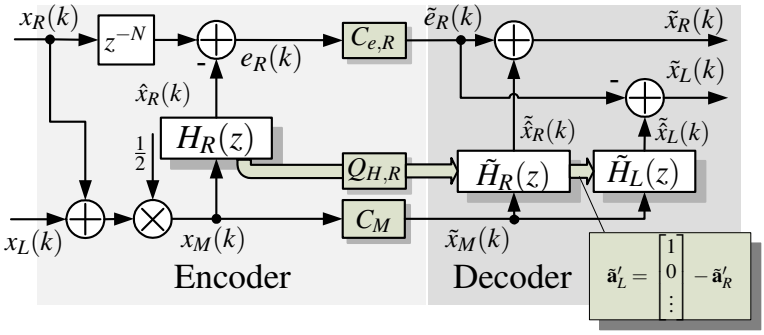

Figure 2 - Simplified coding structure. Here: C_M and $C_{e,R}$ are both AMR-WB

3 Implementation of the Stereo Concept and Use of AMR-WB

To facilitate the application of the stereo extension to a realistic transmission system, certain practical aspects had to be considered compared to [11], where the system was evaluated theoretically for stationary input signals and with a raw quantization model for the transmission of the mono signal and the prediction error. The filter coefficients were assumed to be transmitted error-free and the data rate was distributed freely between the two channels to maximize the sum of logarithmic signal-to-noise ratios (SNR) in the two output channels.

Since AMR-WB (like most current codecs) operates on blocks with a length of 20 ms, this block size was also chosen for the implementation of the proposed stereo extension. The stereo prediction filter coefficients are calculated for $N = 5$ referring to the equations in Section 2.1 for each block with the short-term autocorrelation as an estimate of the autocorrelation function. A rather strong regularization is needed to ensure numerical stability of the matrix inversion in (11). Additionally, applying new filter coefficients in each block leads to filter switching artifacts which can be very annoying for certain signals.A short cross-fade between the filter coefficients of consecutive blocks is introduced to take care of this.

In order to evaluate the capabilities of the linear phase linear prediction based stereo enhancement, any mono core codec could be used for C_M and $C_{e,R/L}$ because the stereo enhancement does not impose any restrictions on the rest of the transmission. One very attractive possibility is AMR-WB, standardized by both ITU ([7]) and 3GPP ([5]). For the evaluation of the stereo enhancement, AMR-WB also has the advantageous property that it offers 9 different data rates. This will be used in the evaluation of the concept to compare different data rate distributions for the mono channel and the prediction error signal. Two individual instances of AMR-WB are used for the mono channel C_M and the prediction error $C_{e,R}$, each of these instances can be set to an individual data rate.

4 Evaluation and Analysis

In the following, the overall stereo coding performance of the new approach will be evaluated by varying the data rates B for both the mono channel (B_M) and the prediction error signal (B_e). The resulting overall SNR will be compared to the performance of AMR-WB for an independent transmission of the two input channels (data rates B_L and B_R for left and right channel respectively. M/S joint-stereo coding will be considered as a benchmark for the achievable data rate reduction in comparison to the individual transmission of left and right channel. The encoding input channels x_L and x_R are the output of a X-Y stereo recording configuration consisting of two cardioid microphones in a shoebox-shaped room. The target of the system analysis is to compute the overall SNR related to the reconstruction of the left and the right channel signals,

$x_L(k)$ and $x_R(k)$ in the decoder, respectively

$$\mathrm{SNR} = 10 \cdot \log_{10} \left(\frac{E\{x_L(k)^2 + x_R(k)^2\}}{E\{(x_L(k) - \tilde{x}_L(k))^2 + (x_R(k) - \tilde{x}_R(k))^2\}} \right), \tag{16}$$

with L/R representing either the left or the right channel. The measured SNR values will be determined for different data rate distributions between mono channel and prediction error.
The simulated recording room has a floor space of 5 by 4 meters and a ceiling height of 3 meters with a reverberation time T_{60} of 150 ms. The omnidirectional sound source S is positioned at a height of 1.9 meters while the microphones are at 1.7 meters. The positioning relative to each other can be seen in figure 3. The microphones are pointing in different directions but they are very close together so that time delay between them can be considered negligible.
This however does not mean that one channel is just an attenuated version of the other because room reflections are picked up at strongly varying sound levels by the two microphones.

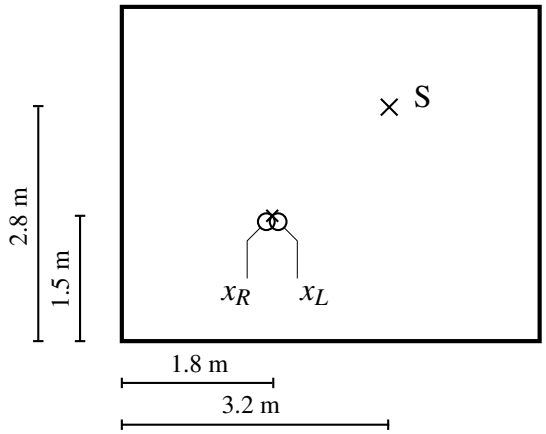

Figure 3 - Room configuration for the generation of the test signals

Test signals were generated by convolving impulse responses (calculated using Allen's and Berkley's image method [1] with the aforementioned room parameters) with mono speech and audio data.

5 Results

For different distributions of the data rate between the channels, measurements of the resulting SNR have been carried out and verified by informal listening tests. Four different test scenarios were considered to assess the performance of the proposed stereo coding scheme:

- Independent transmission of the two input channels
- Transmission of mono channel and prediction error with two instances of AMR-WB as presented in section 3 and transmission of one set of filter coefficients with 1.6 kbit/s
- Parametric stereo - transmission of mono channel with AMR-WB and transmission of one set of filter coefficients with 1.6 kbit/s
- Transmission of sum and difference channel with two instances of AMR-WB referring to the principle of M/S joint-stereo coding

To quantify the expected performance range of AMR-WB for the given input signal the SNR values for an independent transmission of the two channels are listed in Table 1. Both channels are using identical data rates ($B_L = B_R$) in this setup.

B_L and B_R in kbit/s	6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
SNR in dB (16)	2.6	3.4	4.0	4.2	4.3	4.5	4.5	4.6	4.6

Table 1 - SNR values for transmission of two independent signals with the depicted data rate for each channel (total data rate $B = B_L + B_R$)

Using the presented stereo extension with a mono channel data rate B_M of 23.85 kbit/s, a data rate B_F of 1.6 kbit/s for the filter coefficients and varying the data rate B_e for the prediction error leads to the SNR values in Table 2. The stereo extension was designed to concentrate most of the signal energy in the mono channel and the filter coefficients so that this decrease in data rate for the prediction error should not have a significant impact on the quality at the receiver.

B_e in kbit/s	6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
SNR in dB (16)	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6

Table 2 - SNR values for transmission of mono channel with $B_M = 23.85$ kbit/s, filter coefficients with $B_F = 1.6$ kbit/s and prediction error with the depicted data rate B_e (total data rate $B = B_M + B_F + B_e$)

The results are identical for all side channel data rates and at the same level as for a transmission of two independent channels with 23.05 kbit/s each. The next setup to evaluate is the parametric stereo configuration (i.e., only mono channel and prediction filter coefficients are transmitted). For the two highest data rates of AMR-WB, this leads to the same 4.6 dB SNR.

The reference M/S joint-stereo setup also reaches this SNR value but only at significantly higher overall data rates than the presented prediction-based stereo extension.

Comparing the lowest possible data rates for which the SNR is still at 4.6 dB, one finds 46.1 kbit/s for an independent transmission of the signals (23.05 kbit/s for each channel), 35.7 kbit/s for M/S joint-stereo coding (23.05 kbit/s for the mid channel and 12.65 kbit/s for the side channel) and just 24.65 kbit/s for the prediction-based approach (23.05 kbit/s for the mono signal and 1.6 kbit/s for the filter coefficients). Especially the performance for a purely parametric stereo transmission is quite remarkable when considering that M/S joint-stereo without transmission of the side channel only reaches a maximum SNR of 3.9 dB.

6 Conclusion

In this contribution, the application of a linear phase linear prediction approach for joint-stereo coding is proposed. The stereo extension concept was shown to add just very little algorithmic delay and it is well suited for combination with common existing mono core codecs.

In this contribution, AMR-WB was investigated for the transmission of both the mono signal and the prediction error. The coding performance was assessed based on measurements of the achievable SNRs for different overall data rates. All presented coding schemes achieved comparable signal quality at the receiver. However, the presented stereo extension allowed for a reduction of the overall data rate by nearly 50 % compared to an independent transmission of the

signals and by more than 30% compared to M/S joint-stereo coding. The coding performance for the stereo extension almost only depends on the data rate for the mono channel so that even a purely parametric stereo solution can offer a quality at the receiving side comparable to a transmission with two full codecs in the case of independent transmission of the signals. One attractive possibility to transmit the prediction filter coefficients for a parametric stereo extension might be the data hiding scheme from [8].

So far, the extension is only well suited for the transmission of stereo signals with no time delay between the individual channels, e.g recordings with coincident microphone setups. A combination with a delay compensation should allow for a significantly widened range of possible application scenarios.

References

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

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

[3] J. Breebaart and C. Faller. *Spatial Audio Processing*. John Wiley, 2007.

[4] J. Eargle. *Handbook of Recording Engineering*. Springer Science+Business Media Inc., 2002.

[5] ETSI, Rec. TS 26.171. Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description, 2009.

[6] H. Fuchs. Improving Joint Stereo Audio Coding by Adaptive Inter-Channel Prediction. *IEEE Workshop on Applications of Signal Processing to Audio and Acosutics*, 1993.

[7] 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)*, volume IV, pages 533–536, Honolulu, Hawai'i, USA, Apr. 2007.

[8] J. Herre, K. Brandenburg, and D. Lederer. Intensity Stereo Coding. *AES 96th Conv.*, pages 1–10, 1994.

[9] http://www.answers.com/topic/fm broadcasting. FM broadcasting, 2007.

[10] ITU-T Rec. G.722.2. Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), 2003.

[11] H. Krüger and P. Vary. A new approach for low-delay joint-stereo coding. In *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, Oct. 2008.

[12] L. Lajmi. *Paketsubstitution in Audiosignalen bei paketorientierter Audioübertragung*. PhD thesis, Technische Universität Berlin, Apr. 2003.

[13] T. Liebchen. Lossless Audio Coding Using Adaptive Multichannel Prediction. *113th Conv. of the Audio Eng. Soc.(AES), Los Angeles, USA*, 2002.

[14] E. L. Torick and T. B. Keller. Improving the signal to noise ratio and coverage of FM stereo broadcasts. *AES Journal*, 33(12), dec 1985.