MDCT-BASED CODER FOR HIGHLY ADAPTIVE SPEECH AND AUDIO CODING

Guillaume Fuchs, Markus Multrus, Max Neuendorf and Ralf Geiger

Fraunhofer Institut Integrierte Schaltungen
Am Wolfsmantel 33, 91058, Erlangen, Germany
email: amm-info@iis.fraunhofer.de

ABSTRACT

Coding audio material at low bit rates with a consistent quality over a wide range of signals is a current and challenging problem. The high-granularity switched speech and audio coder AMR-WB+ performs especially well for speech and mixed content by promptly adapting its coding model scheme to the signal. However, the high adaptation rate is done at the price of limited performance for non-speech signals. The aim of the paper is to enhance the coding efficiency of AMR-WB+ while maintaining its high flexibility. For this purpose, the original DFT was replaced by the state-of-art transformation MDCT, and the vector quantization by the combination of a scalar quantization and an evolved context-adaptive arithmetic coder. The improvements were measured by both objective and subjective evaluations.

1. INTRODUCTION

Designing a low bit-rate audio coder performing well independent of the content has been a recurrent topic of research for the last years. Historically, low bit-rate speech and audio codecs have been developed for different target applications and with different assumptions. As a result, state-of-the-art coding schemes show unbalanced quality dependending on the signal nature, especially at low bit-rates. Speech coders usually rely on a speech production model, like CELP, while generic audio coders use a transformation of the signal over relatively long analysis windows.

AMR-WB+ [1] provides one attempt to unify speech and audio coding paradigms, which is specifically well adapted for speech and mixed content audio materials. It is based on a high granularity switch between a speech core-coder, ACELP (*Algebraic Code Excited Linear Prediction*), and transform-based core-coders, TCX (*Transform Coded eXcitation*), working with different time/frequency resolutions. The appropriate selection of the core-coding modes permits AMR-WB+ to achieve good performance for clean speech and mixed content at low bit-rates [2]. However, the high adaptation rate of AMR-WB+ makes it difficult to adopt state-of-art generic audio coding tools and optimizations. For instance, it is constrained to use small overlaps. As a result, more stationary audio signals, like music, are better handled by standalone transform-based coders, like MPEG-4 High Efficiency AAC [3]. For music-like signals, the coding should ideally be switched at a lower granular scale to a dedicated generic audio core-coder, adopting longer overlapping regions and more evolved perceptual cues. This global approach was already proposed in [4].

The objective of this paper is to improve the AMR-WB+ coding efficiency for the monophonic case and especially the TCX coding. The TCX coding originally proposed in [5] was based on a Discrete Fourier Transform (DFT). The coding scheme was studied in several works and improvements were already proposed when using it as a standalone coder with no signal adaptation [6, 7] or when using only window length adaptation [8].

However within a highly adaptive switched coder like AMR-WB+, conciliating high efficiency and high adaptation is more delicate. The main contribution of this paper is to design an efficient TCX coding for AMR-WB+ while maintening its high adaptation capability. The DFT-based TCX within AMR-WB+ suffers from some limitations, in particular during the core-coding transitions where TCX produces a significant overhead information and rather abrupt changeovers. For this purpose, the paper proposes to replace the original transformation block by an MDCT. The aims are to go toward critical sampling to improve the frequency response and to get smoother transitions. As a consequence several TCX tools must be adapted and optimized for the new transformation. Furthermore, the original Lattice Vector Quantization (LVQ) was replaced advantageously by a scalar quantization followed by a context-adaptive entropy coder using appropriate resamplings of the context when switching from one core-coding mode to another.

2. THE STANDARD AMR-WB+

The 3GPP AMR-WB+ standard is an extension of the AMR-WB speech coder [9], adding a selectable frequency domain coding, parametric bandwidth extension, and parametric stereo coding. In this way, the capability of coding music and speech over music was improved significantly. The low-frequency mono signal is encoded using either a speech coder or a transform-based coder. The speech coder is based on ACELP, while TCX with a Lattice Vector Quantization (LVQ) is used for the transform-based coding. The signal is processed on a super-frame basis of 1024 samples. Each super-frame is decomposed into 4 frames of non-overlapped 256 samples. Any frame can be encoded using either ACELP or TCX, while the first two and the last two frames within a super-frame can be encoded instead by a single medium size TCX, and the whole super-frame by a single long size TCX. Subsequently, the different TCX modes are called TCX-256, TCX-512 and TCX-1024. In total 26 different mode combinations within a super-frame are possible.

The mode selection can be done in either an open-loop or a closed-loop fashion. The open-loop decisions are taken by a previous analysis of the signal and are based on audio signal characteristics. The closed-loop decisions are made by comparing in an efficient way all possible combinations after the whole encoding-decoding process. The best combination is selected according to the segmental Signal to Noise Ratio (segSNR) calculated in a weighted domain, which is more relevant for the perceptual quality than in the signal domain.

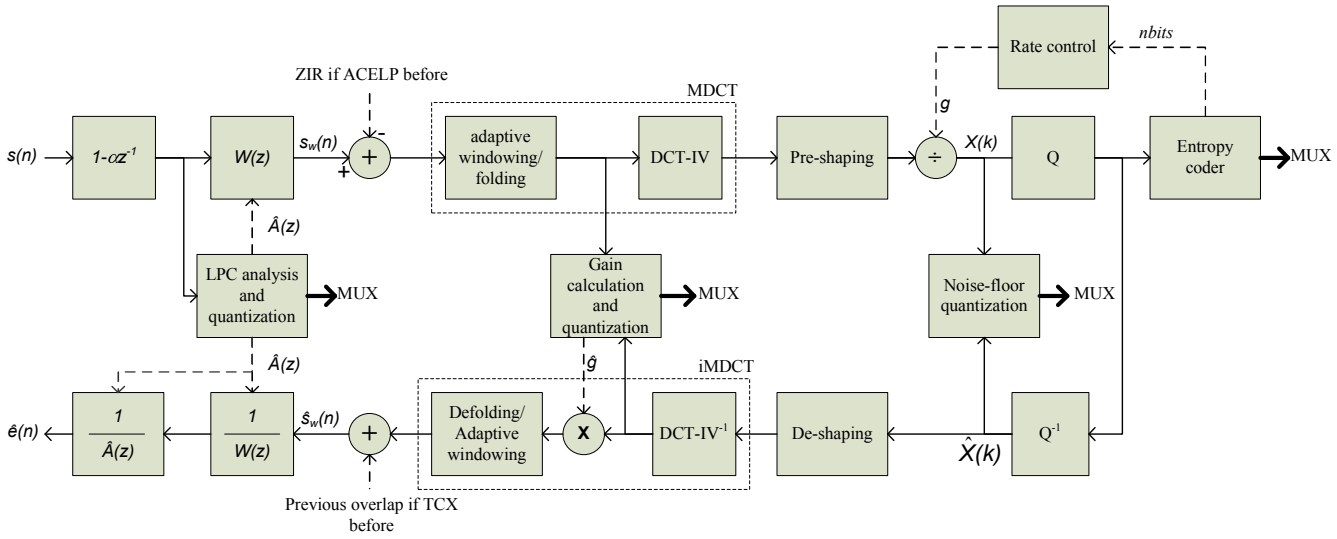

Figure 1: MDCT-based TCX encoding principle

The closed-loop decision shows very reliable results and is an important feature of the AMR-WB+. The closed-loop decisions will be used in the rest of the paper.

AMR-WB+ can accept different input sampling rates from 12.8 to 38.4 kHz, and can deliver bit-rates lower than 10 kbps up to 48 kbps. Nevertheless, the nominal sampling rate is 25.6 kHz and the usual working bit-rate is between 12 and 32 kbps. The core-coders are fed with the input signal downsampled by a factor of 2 representing the low frequency band, while the high frequency band is conveyed by the bandwidth extension. The core-coder works usually between 10 kbps and 30 kbps.

3. ENHANCED AMR-WB+

The paper proposes to modify the AMR-WB+ by enhancing the transform-based core-coder TCX. The paper does neither consider the bandwidth extension nor the parametric stereo, knowing that both of these AMR-WB+ features can be subject of improvements as well.

3.1 MDCT-based TCX principle

Our approach is based on replacing the original DFT in the TCX core-coder by an MDCT. Subsequently the new TCX coding is named MDCT-based TCX. The block diagram of the MDCT-based TCX encoding is shown in Figure 1. As in the original AMR-WB+, the signal is locally decoded at the encoder side for letting the closed-loop decisions compute the segSNR on $\hat{s}_w(n)$ and for feeding the future ACELP adaptive codebook with the synthesized excitation $\hat{e}(n)$.

As in conventional TCX, the MDCT-based TCX encodes a target signal output of a pre-emphasis filter $1-\alpha z^{-1}$ and a weighting filter $W(z)$ defined as:

$$W(z) = \frac{\hat{A}(z/\gamma)}{1-\alpha z^{-1}}$$

where $\alpha = 0.68$ is defined as the tilt factor and $\gamma = 0.92$. The LPC coefficients $\hat{A}$ are issued from a linear prediction analysis of order 16, quantized in the Immittance Spectral Pair (ISP) domain and finally interpolated in the same domain to get a set of parameters every 64 samples.

The Zero Impulse Response (ZIR) is subtracted from the target when ACELP is used as core-coder in the previous frame. This subtraction helps the frequency analysis by fading in the signal to transform. The window adaptation is a function of the current core-coding mode and the previous mode. The window sequence is explained in detail in the next section. The MDCT is then applied by means of a folding and a DCT-IV. The low frequencies are pre-shaped as it is done in AMR-WB+. The quantization is based on a gain-shape decomposition, where the global gain g is applied to each frequency coefficient. g defines the uniform scalar quantization step and at the same time the bit allocation. Indeed, the higher g, the coarser is the quantization and the lower is the bit demand. Based on the calculation of the consumed bits, *nbits*, by the entropy coder, g is adjusted in order to match the consumption to the bit budget as closely as possible.

The zeroed frequency coefficients serve to calculate a noise floor which is injected at the decoder as described later. At the encoder, the signal is locally reconstructed in the weighted domain, where the segSNR is computed and used for the closed-loop decision. During the synthesis, the global gain is recalculated and quantized as explained later on. The common overlapping region from the previous TCX window is added to cancel the time domain aliasing components introduced by the forward MDCT.

3.2 Window sequence

TCX uses low overlapping windows which are schematized in Figure 2. The windows are composed of three parts, a left overlap of L samples, a middle region of M non-overlapped samples, and a right overlap of R samples. The overlapping regions of the windows consist of sine slopes, while the middle region is windowed by ones.

In the original AMR-WB+, R is fixed to be 1/8th of $L+M$, where $L+M$ covers an integer number of frames. The first $L+M$ samples of the current window can be recon-

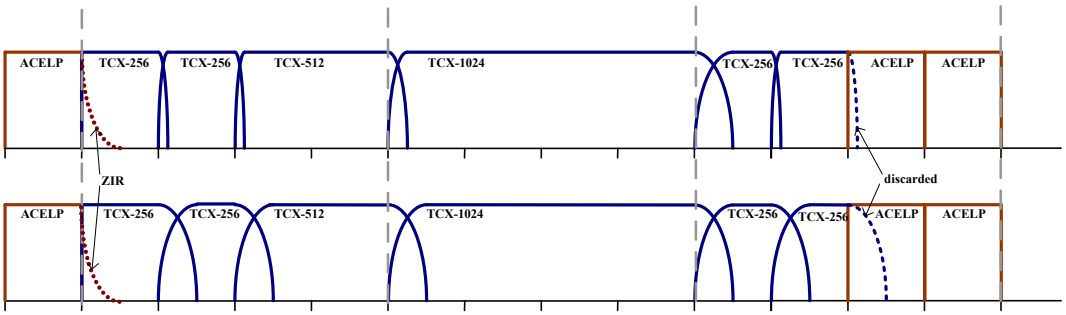

Figure 3: Comparison between the DFT-based TCX (top) and MDCT-based TCX (bottom) window sequence

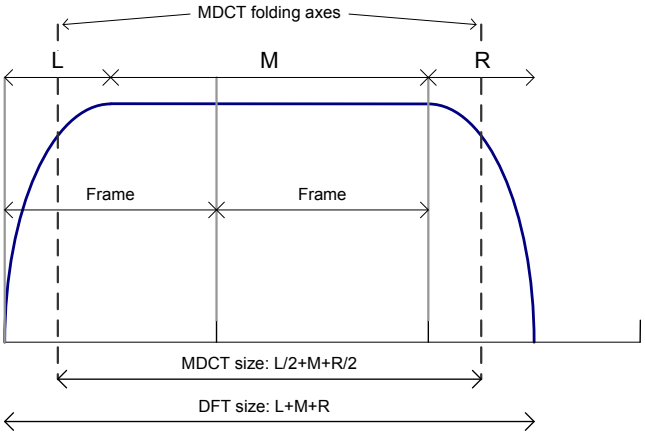

Figure 2: TCX window shape for the 512 samples case. The MDCT size refers to the number of transformed coefficients and is reduced compared to the original DFT size.

structed using the right overlap from the previous window. For this purpose, L is equal to the previous right overlap size R when the last coding was TCX, or set to 0 when coming from ACELP, since ACELP uses implicit rectangular windows. In the latter case, the transition is smoothed by the ZIR subtraction. Thus, the number of DFT coefficients is equal to $N = L+M+R$, and the overhead information represents 1/8th of the global bit-rate when using TCX. Therefore, the overlaps are kept reasonably small.

Replacing DFT by MDCT permits to reduce the overhead information. The number of transformed coefficients is reduced to $N = L/2+M+R/2$. The transitions between two consecutive TCXs are now critically sampled. The only remaining non-critical sampling occurs during transitions from TCX to ACELP, where the $R/2$ folded samples of the right overlap are discarded.

The new window sequence is compared to the old one in an example given at Figure 3. In the MDCT case, the critical sampling allows for larger and more homogeneous transitions. For each size of MDCT-based TCXs, R and L are fixed to 128 except when coming from ACELP where L is still set to zero. As a consequence, the new transitions are now better smoothed, and the enlarged analysis windows enhance the energy compaction and consequently, the coding performance.

3.3 Entropy-coded scalar quantization

The quantized, scaled and transformed coefficients are coded by a context-adaptive arithmetic coder. The coefficients are gathered in 4-tuples, and each 4-tuple is decomposed further into bit planes. The two most significant signed bit planes form a symbol which is encoded using probabilities derived from the surrounding context as proposed in [10]. The least significant planes use a uniform probability distribution assumption. The symbol coming from the two most significant signed bit planes and the remaining planes are fed into an arithmetic coder with their respective probabilities.

The context is calculated synchronously at both encoder and decoder sides using past transmitted information as illustrated in Figure 4. In the implemented version, the context considers four already coded 4-tuples in the direct neighbourhood of the current 4-tuple. The derived context is then mapped to one of the 32 probability models generated during a training phase. Further, the probability models are efficiently stored in memory by quantizing the symbols from the two most significant signed bit planes into 544 groups. Inside a group, the probability is assumed to be similar. Finally, only 544 cumulative frequencies for each model need to the stored.

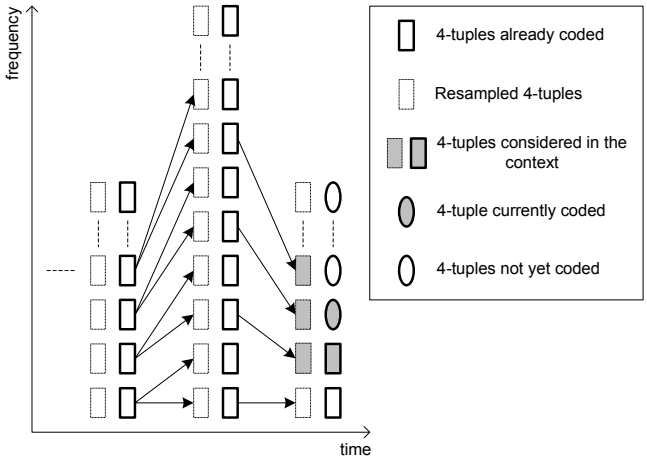

Figure 4: Illustration of the context resampling and calculation used in the entropy coder

The context adaptation is a highly efficient approach, but not trivial to apply directly to the AMR-WB+ structure. In-

deed, the 4-tuples considered in the context calculation can come from a past TCX with different time/frequency resolution than the current TCX. As shown in Figure 4, a resampling is then necessary for changing the frequency resolution of the transmitted coefficients in the past TCX into the frequency resolution of the current TCX. The resampled coefficients $\hat{X}_r(k)$ are obtained as follows:

$$\hat{X}_r(k) = \hat{X}_p\big(4 \cdot \lfloor \lfloor k/4 \rfloor \cdot ratio \rfloor + \mathrm{mod}(k,4)\big)$$

where $\hat{X}_p$ are the N_p coefficients of the past TCX, $k = \{0..N-1\}$ is the index of the current transformed coefficients, $\mathrm{mod}(k,4)$ is the integer remainder of the division of k by 4, and *ratio* is defined as:

$$ratio = \frac{N_p}{N}$$

3.4 Noise filling

The noise filling plays an important role in the AMR-WB+. Because the different modes share the same core bandwidth, TCX modes must deal with the large bandwidth imposed by ACELP. For low bit-rates, transform-based coders usually handle narrower bandwidths. The noise filling permits to mask the artifact coming from the zeros introduced in the quantization process by filling them with an artificial noise. Originally, AMR-WB+ uses a vector quantization which processes the spectrum by vectors of dimension 8. This means that the noise filling is applied only when 8 samples within a vector are zeroed. The noise filling associated with the scalar quantization tries to mimic the behaviour by considering only long runs of zeros. It avoids concealing non-zero quantized values by the injected surrounding noise. The runs of zeros are detected as follows:

$$r_0(k) = \begin{cases} 0 & \text{if } k < N/6 \\ 0 & \text{if } \sum_{i=0}^{7} |\hat{X}(8 \cdot \lfloor k/8 \rfloor + i)|^2 > 0 \text{ and } N/6 \le k < N \\ 1 & \text{otherwise} \end{cases}$$

The noise floor nf is calculated as the root mean square of the zeroed values before being quantized to 3 bits on a logarithmic scale going from 0 to 0.5.

$$nf = \sqrt{\frac{\sum_{k=0}^{N-1} (X(k) \cdot r_0(k))^2}{\sum_{k=0}^{N-1} r_0(k)}}$$

The injected noise $R(k)$ is a white noise generated by randomly changing the sign of the unity. The noise filling is then applied to the dequantized values as follows:

$$\hat{X}'(k) = \begin{cases} \hat{X}(k) & \text{if } r_0(k) = 0 \\ R(k) \cdot \widehat{nf} & \text{otherwise} \end{cases}$$

where $\widehat{nf}$ is the dequantized noise floor.

3.5 Gain calculation

In AMR-WB+, the global gain g, used in the gain-shape quantization, is not directly transmitted but recomputed in time domain by means of a scalar product between the windowed original signal and the windowed synthesis. This is done to emphasize the waveform preservation and, as a matter of fact, maximize the segSNR criterion used further for the closed-loop decision. In MDCT-based TCX, the reconstructed time signal after the inverse transformation includes time domain aliasing components. The gain should then be calculated between the windowed and folded original signal and windowed and folded reconstructed signal as it is specified in Figure 1. Thus, time domain aliasing components and overlapping regions are taken into account in the gain calculation. The gain is then quantized on a logarithmic scale to 7 bits.

4. CODING PERFORMANCE

4.1 Objective evaluation

The segSNR in the weighted domain was used for assessing the improvements brought to AMR-WB+ as it is the criterion on which relies the closed-loop decision. For the evaluation, the segSNR was averaged over a long audio sequence containing speech, mixed content and music. Three different flavors of TCX were compared: the original TCX combining DFT and LVQ (DFT-LVQ), a hybrid TCX combining MDCT and LVQ (MDCT-LVQ), and the proposed MDCT-based TCX combining MDCT and the entropy-coded scalar quantization (MDCT-SQ).

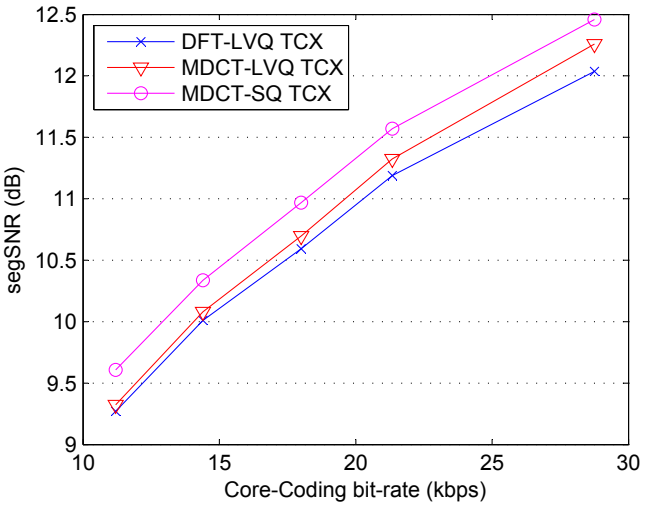

Figure 5: Segmental SNR Performance of AMR-WB+ core-coding using different flavours of TCX

The performance results are plotted in Figure 5 for core-coding bit-rates going from 11.2 to 28.75 kbps. One can observe the MDCT improvement by comparing DFT-LVQ and MDCT-LVQ. Besides, the contribution of the entropy-coded scalar quantization is seen by comparing MDCT-LVQ and MDCT-SQ. As a general observation, the improvement is maintained over the whole bit-rate range. Moreover, the gains increase with the bit-rate especially the one coming from MDCT. This behaviour can be explained by the fact that the TCX modes are more often used as the bit-rate increases. Consequently, the benefit of MDCT over DFT becomes more obvious. Finally, the entropy-scaled quantization performs significantly better than LVQ for all bit-rates. It is worth noting that the average segSNR takes into account ACELP coded frames which do not benefit from the improvement.

Table 1 details the repartition of the gain of the enhanced AMR-WB+ amongst the different TCX modes. At low bit-rate, the longest TCX, TCX-1024, benefits most from the enhancements, while at high bit-rate, they are the most profitable for the shortest TCX, TCX-256.

Bit-rate	TCX-256	TCX-512	TCX-1024
14 kbps	0.25	0.35	0.35
29 kbps	1.02	0.90	0.75

Table 1: SegSNR gains for the different TCX modes of MDCT-based TCX over DFT-based TCX expressed in dB

4.2 Subjective evaluation

A listening test was conducted in order to evaluate the perceptual quality of the improvements. The enhanced AMR-WB+ using the MDCT-based TCX was compared to AMR-WB+ using the original DFT-based TCX. Both versions used the same bandwidth extension and the same ACELP speech coder. The test was done for an overall bit-rate of 20 kbps.

Eight expert listeners participated in the test. Ratings were given within a paired comparison blinded test according to [11] on a 7-points comparison scale form -3 to 3. Positive scores are in favour of the enhanced AMR-WB+. The test comprised the following items: two clean speech items, two music items, two speech between music items (SbM) and two speech over music items (SoM).

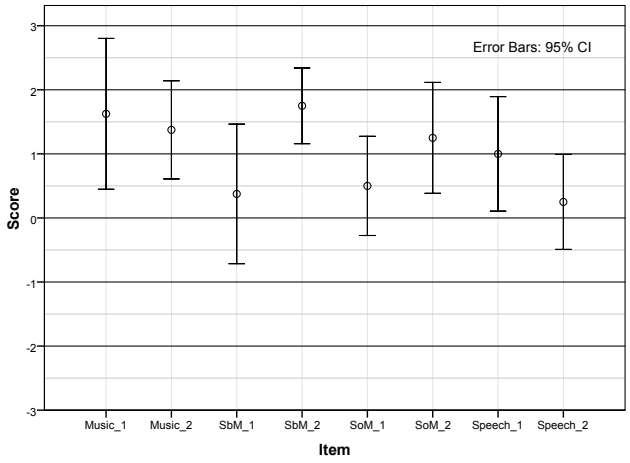

Figure 6: Listening test results at 20 kbps. The enhanced AMR-WB+ version outperforms significantly AMR-WB+ in 5 out 8 items

Figure 6 plots the results of the evaluation. The enhanced AMR-WB+ was statistically preferred for more than half of the items. Moreover, none of the items were determined as worse. As expected, the improvements are mainly detectable where a high presence of music occurs. On clean speech, the TCX modes are seldom selected and the potential improvements less noticeable.

5. CONCLUSION

In this paper, we proposed to improve the AMR-WB+ by replacing the transformation and the quantization process. The modifications maintain the flexible adaptation of the AMR-WB+. The closed-loop decision principle was kept unchanged and no additional delay was added. The modifications were proven to enhance the coding performance by objective and subjective evaluations. The enhancements are especially beneficial for music components of mixed content. The proposed MDCT-based TCX is also used in a larger switched coding scheme which successfully unifies speech and audio coding [4].

Acknowledgment

The authors would like to thank Bruno Bessette, Philippe Gournay, Roch Lefebvre and Redwan Salami from VoiceAge Corporation for their fruitful insight into the details of the AMR-WB+ standard and TCX coding.

REFERENCES

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

[2] J. Mäkinen et al., "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, March 2005, vol. 2, pp. 1109–1112.

[3] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," in *115th AES Convention*, New York, NY, USA, Oct. 2003, preprint 5871.

[4] Max Neuendorf et al., "Unified speech and audio coding scheme for high quality at low bitrates," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2009.

[5] R. Lefebvre, R. Salami, C. Laflamme, and J.-P. Adoul, "8 kbit/s coding of speech with 6ms frame-length," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Oct. 1993, vol. II, pp. 612–615.

[6] J.-H. Chen and D. Wang, "Transform predictive coding of wideband speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 275–278.

[7] M. Oger, S. Ragot, and M. Antonini, "Transform audio coding with arithmetic-coded scalar quantization and model-based bit allocation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Honolulu, Hawaii, USA, april 2007.

[8] Sean A. Ramprashad, "The Multimode Transform Predictive Coding Paradigm," *IEEE Trans. on Speech and Audio Processing*.

[9] B. Bessette et al., "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[10] Nikolaus Meine and Bernd Edler, "Improved quantization and lossless coding for subband audio coding," in *118th AES Convention*, Barcelona, Spain, May 2005, Preprint 6468.

[11] International Telecommunication Union, "General methods for the subjective assessment of sound quality," 2003, ITU-R, Recommendation BS. 1284-1.