# IMPROVED ARITHMETIC CODING FOR TIME-WARPED MDCT BASED AUDIO CODING

*Stefan Bayer and Bernd Edler*

International Audio Laboratories Erlangen
Am Wolfsmantel 33, 91058 Erlangen, Germany
email: stefan.bayer@audiolabs-erlangen.de

**ABSTRACT**

The Time-Warped Modified Discrete Cosine Transform (TW-MDCT) improves the energy compaction for harmonic signals with varying fundamental frequency compared to the plain MDCT. Adaptive context based entropy coding has the potential to provide higher gain over memoryless entropy coding. But in combination with the TW-MDCT, the context based adaptive coding may lead to suboptimal coding. This paper presents an algorithm for improving the context for the TW-MDCT. This is mainly achieved by exploiting already available information on the frequency variation needed by the TW-MDCT. This results in an improved entropy coding.

***Index Terms***— Audio Coding, Entropy Coding, TW-MDCT, Arithmetic Coding

## 1. INTRODUCTION

The Modified Discrete Cosine Transform (MDCT) is widely used in audio coding. Recently, the Time-Warped MDCT [1] was introduced to improve the coding gain for harmonic signals with varying fundamental frequency. The TW-MDCT removes the smearing of higher harmonics occurring in the MDCT for such signals by reducing or completely removing the frequency variation within one processed frame. This results in better energy compaction in the spectrum, or in other words, leads to spectra with a nice harmonic structure instead of smeared, more noise-like looking spectra and therefore to an increased coding gain for such signals.

Context adapted arithmetic coding exploits a higher order entropy by going from a memoryless source model to a model including memory. This can be done by taking into account the context, that is a state derived from already coded past values into account. This principle in combination with arithmetic coding was already used in modern video coders [2] and also proposed for audio coding [3].

Both the TW-MDCT and a context adaptive arithmetic entropy coding scheme [4] for coding the quantized spectral data were incorporated into the new MPEG Unified Speech and Audio Coding (USAC) standard [5, 6]. The coding scheme groups the spectral coefficients into blocks of 2, called 2-tuples. The adaptive context itself is associated with already coded 2-tuples, both neighboring the 2-tuple to code in time and frequency as shown in Figure 1.

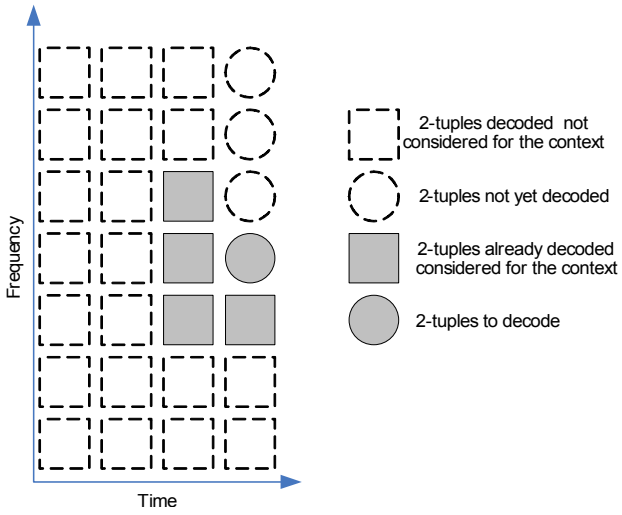


**Fig. 1**. Time-frequency representation of the context

Looking at Figure 2, which shows the unquantized and quantized spectra of a synthetic plucked string with a varying fundamental frequency for two consecutive frames, the TW-MDCT results in harmonic spectra with low smearing. The harmonic grid itself is determined by the average fundamental frequency within one block. One can clearly see that the two consecutive TW-MDCT spectra approximately are the same, only stretched or shrunk along the frequency axis.

For context adaptive arithmetic coding, the context is derived mainly from the past quantized values. When looking at a situation as in Figure 2b, it comes quite naturally that the context will be suboptimal where the harmonic lines no longer occur at the same bins as in the last frame. Looking at the spectral bit distribution for the above case confirms this, as can be seen in Figure 3, showing the same quantized spectra as before, only zoomed in at the frequency bin axis, and

The International Audio Laboratories Erlangen is a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS, Germany.

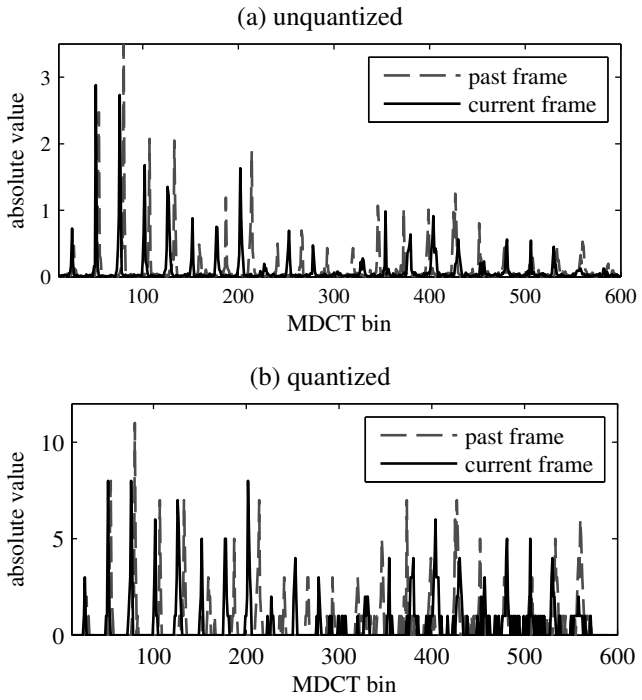


**Fig. 2**. TW-MDCT spectra of two consecutive blocks of a synthetic plucked string with varying fundamental frequency

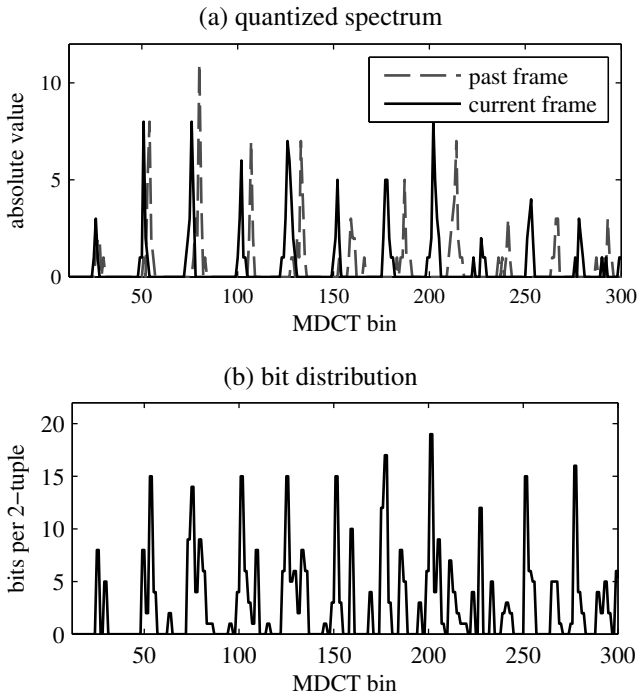


**Fig. 3**. Quantized spectrum, unmapped context and bit distribution for the arithmetic coding (zoomed in)

the corresponding spectral bit distribution when using the unmodified arithmetic coder. Where the positions of the individual harmonics still approximately match, as it is the case for the first partial tone, the coding still works efficiently. On the other hand, for bins now quantized to zero where there was a harmonic line in the last frame and vice versa, the bit demand increases. In most of these cases, the encoder then would determine that it is more efficient to use a context reset to a basic state instead of using the suboptimal past context. In the present work, a new mapping algorithm for the context is proposed, based on the already available frequency variation needed for the TW-MDCT algorithm. The mapping factor is derived from the frequency variation and two different algorithms for the mapping are proposed, one a mapping of 2-tuples and one a mapping of individual lines plus new calculation of the context. An evaluation of both algorithms show that they outperform the arithmetic coder without time-warped context mapping.

## 2. CONTEXT ADAPTATION

### 2.1. Principle

As shown above, when two consecutive blocks of a harmonic signal are processed by the TW-MDCT, they can exhibit a great similarity except for a dilation along the frequency axis. The basic idea behind the warped context adaptation is to stretch the context derived from the past quantized spectral values accordingly. This needs two steps, first deriving the mapping factor from the available frequency variation estimation, and then mapping the past context based on this factor. Additionally, a flag is introduced to the bit stream to decide to use the adapted context only for frames where it yields a better compression than the unmapped context.

### 2.2. Mapping Factor

In the TW-MDCT the reduction of frequency modulation is done by a time-varying resampling. The resampling is carried out so that for a given frame the number of samples in the warped and the linear time are equal. This means that the signal in frame $k$ is now represented by a harmonic spectrum where the fundamental frequency $\bar{F}_{0,k}$ is the average of the varying fundamental frequency $F_0[n]$ in linear time over the frame:

$$\bar{F}_{0,k} = \frac{1}{N} \sum_{n=kN}^{(k+1)N-1} F_0[n] \tag{1}$$

The mapping factor for the warped context mapping then simply would be the ratio of the two consecutive average fundamental frequencies:

$$r_k = \frac{\bar{F}_{0,k-1}}{\bar{F}_{0,k}}. \tag{2}$$

In the recently proposed TW-MDCT algorithm [1], the control information for the time warping is not the absolute fundamental frequency, but only the relative frequency variation information $f_{rel}$, since this is sufficient. Because the stretch factor is derived by the ratio of the two (absolute) average fundamental frequencies it is equal to calculating the average relative fundamental frequencies of two consecutive frames based on a smoothed concatenation of the relative frequency curve of two consecutive frames:

$$r_k = \frac{\sum_{k-1} f_{rel}}{\sum_k f_{rel}} \tag{3}$$

Both sums in the equation are already available from the TW-MDCT algorithm and need not to be calculated specifically for the warped context mapping.

### 2.3. Algorithm I: Tuple-wise mapping

For the first algorithm, the contexts $c[n]$ for the 2-tuples are simply mapped tuple-wise:

$$c_{w,k}[n] = c_k[\lfloor nr_k \rceil], 1 < n \leq min\left(n_{max}, \lfloor n_{max} r_k \rceil\right) \tag{4}$$

where $\lfloor \cdot \rceil$ denotes the rounding to the nearest integer and $n_{max}$ is the maximum number of 2-tuples used for coding. If the number of tuples after warped context mapping is smaller than $n_{max}$, the remaining contexts are set to a predetermined state $c_{w,k}[n] = 1$.

### 2.4. Algorithm II: line-wise mapping

In the considered arithmetic coding scheme [4], the context $c_k$ of a 2-tuple is determined by the absolute value of the quantized spectral lines:

$$c_k[n] = min\left(|a_k[n]| + |b_k[n]| + 1, 15\right) \tag{5}$$

where $a_k$ and $b_k$ are the quantized spectral lines building a 2-tuple. The second algorithm first rebuilds the past quantized spectral data.

$$q_k[2n-1] = a_k[n], q_k[2n] = b_k[n], 1 < n \leq n_{max} \tag{6}$$

Then the spectral lines are mapped according to the stretching factor:

$$q_{w,k}[m] = q_k[\lfloor mr_k \rceil], 1 < m \leq min\left(m_{max}, \lfloor m_{max} r_k, \rceil\right) \tag{7}$$

where $m_{max} = 2n_{max}$ is the maximum number of spectral lines used for coding. Similar to the first algorithm, if the number of lines after warped line mapping is smaller than $m_{max}$, the remaining lines are set to zero. Finally the spectral lines again are grouped into 2-tuples

$$a_{w,k}[n] = q_{w,k}[2n-1], b_{w,k} = q_{w,k}[n], 1 < n \leq n_{max} \tag{8}$$

and using equation 5 the new past context is calculated. The adaptation for schemes with different tuple sizes is straightforward.

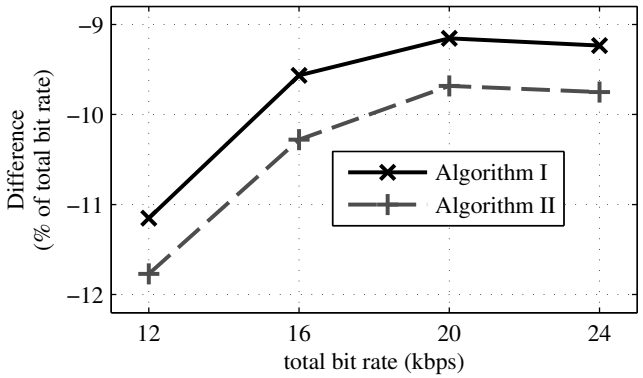


**Fig. 4**. Relative reduction in bit consumption using context mapping in comparison to unmapped context for the idealized case

## 3. EVALUATION

For the evaluation of the proposed algorithms, first USAC compliant bit streams were generated without using the time-warped mapping. Then the decoded quantized spectral values were again encoded using the two different proposed algorithms and the potential savings calculated. The arithmetic coder itself was not changed, meaning no training of the context mapping and context clustering according to [4] was done for the time-warped contexts.

The gain of the proposed algorithm depends on the accuracy of the frequency variation in the encoder. To get a sense of the overall achievable gain, a synthetic harmonic signal with an exactly known moderate frequency modulation was encoded using the frequency variation information in an uncoded manner. Figure 4 shows that in such an ideal case the difference is quite significant.

For a realistic comparison, a set of real world items and an encoder employing an estimation algorithm for the frequency variation with subsequent coding was used. The test set consists of the same items used for the evaluation of the proposals during the standardization process of USAC, containing a mixture of music, speech over music and clean speech items. Figure 5 shows the achieved improvement for different bit rates. The savings are around 0.5% of the total bit rate used for coding at the low end of the used bit rate range and decrease towards higher total bit rates. This decrease comes from the increasing sampling frequency used for higher bit rates and therefore for a constant MDCT transform length the time resolution increases and the frequency resolution decreases. At one hand, this reduces the possible frequency variation within one frame and therefore less gain for the TW-MDCT itself compared to the plain MDCT. Also the mapping factors are smaller than for lower sampling frequencies and therefore there is less mismatch between unmapped and mapped contexts.

Furthermore the savings are of course higher when only

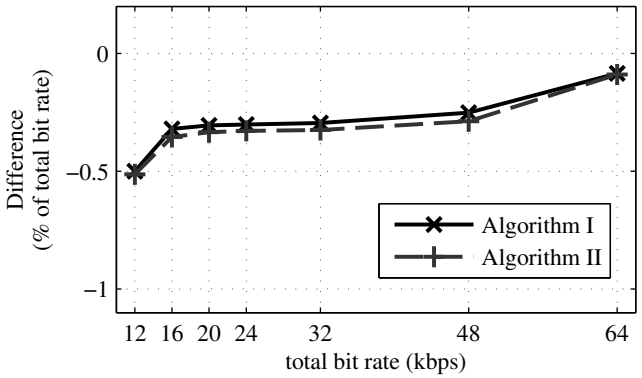


**Fig. 5**. Relative reduction in bit consumption using context mapping in comparison to unmapped context for the real world items

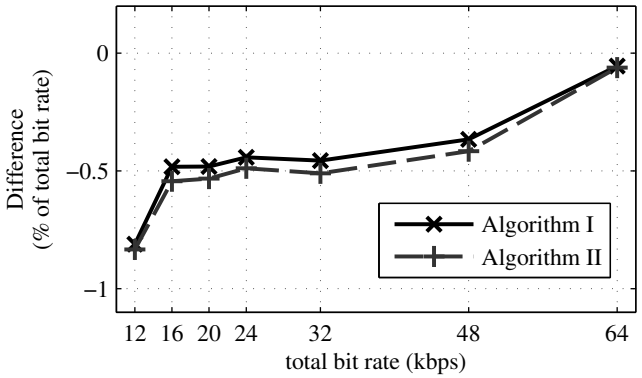


**Fig. 6**. Relative reduction in bit consumption using context mapping in comparison to unmapped context for the real world items (only for frames where mapping is possible).

considering frames where a mapping is possible, that is where the mapping factor is large enough so that at least one tuple or line gets mapped. This can be seen in figure 6. That also means that the overall gain depends on the signal characteristics, signals where there are frequency modulated signal portions will gain more from the proposed algorithms than more noise-like signals.

Another observation is that the line-wise mapping yields better results than the tuple-wise mapping, due to the finer frequency resolution for the line-wise mapping. For arithmetic coding schemes that take even more lines into account for one tuple, this advantage should increase.

One further look is at the context reset mechanism. The encoder also uses a default context where the context $c[n]$ is set to a predetermined state for all 2-tuples . The encoder both encodes with the context derived from the actual past frame and the default context, and if the default context gives better compression, a reset flag is sent to the decoder. Using the time-warped mapping now yields a much higher number

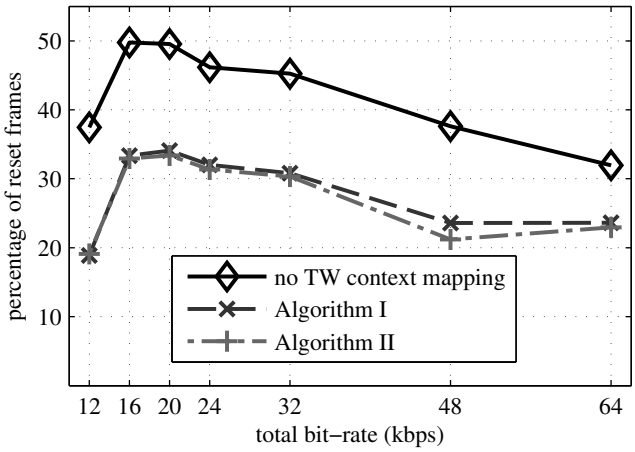


**Fig. 7**. Percentage of reset frames where TW context mapping is possible

of frames where the mapped context is better suited than the default context. This reduces the number of frames where the context is reset, as can be seen in figure 7, where the amount of reset frames drops by approximately 15%.

The algorithms itself are relatively cheap in terms of computational complexity compared to overall complexity of the arithmetic coder and no additional memory is needed for the operations. The mapping flag only needs to be added to the bit streams in cases where there is no reset and where mapping is possible due to time warping, leading to less than 1 bit per frame for this additional side information on average.

## 4. CONCLUSION

This work provides an improvement to adaptive context arithmetic coding [2–4] when used in combination with a TW-MDCT filter bank [1] within an audio coder [5, 6]. A new context mapping algorithm for an context adaptive arithmetic coder used for coding TW-MDCT spectra is presented. First, the contexts are suboptimal for a TW-MDCT based transform coder. This was followed by presenting a new algorithm for generating better suited past contexts when time-warping is active. Two different approaches were proposed, one with tuple-wise, the other with line-wise mapping. For evaluation first an idealized example was presented to show the potential of the algorithm. Then a realistic comparison was made showing a slight overall improvement of the coding efficiency, which is larger when only looking at frames where a mapping is possible. Also, the number of frames increases where the past context is now better suited than a default context. Possible further improvements could include training of the arithmetic coder context mapping and context clustering based on the proposed time-warped mapping. The results for the idealized case imply that improvements in the estimation and coding of the frequency variation could lead to larger savings.

# References


[1] Bernd Edler, Sascha Disch, Stefan Bayer, Fuchs Guillaume, and Ralf Geiger, "A Time-Warped MDCT Approach to Speech Transform Coding," in *126th AES Convention*, Munich, Germany, May 2009, Preprint 7710.

[2] Detlev Marpe, Heiko Schwarz, and Thomas Wiegand, "Context-Based Adaptive Binary Arithmetic Coding in the H.264/AVC Video Compression Standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 620 –636, jul. 2003.

[3] Nikolaus Meine and Bernd Edler, "Improved Quantization and Lossless Coding for Subband Audio Coding," in *118th AES Convention*, Barcelona, Spain, May 2005, Preprint 6468.

[4] Guillaume Fuchs, Vignesh Subbaraman, and Markus Multrus, "Efficient Context Adaptive Entropy Coding for Real-Time Applications," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2011, vol. I, pp. 493–496.

[5] Max Neuendorf et al., "MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of All Content Types," in *132nd AES Convention*, Budapest, Hungary, April 2012, Preprint 8654.

[6] ISO/IEC International Standard 23003-3:2012, "MPEG-D (MPEG Audio Technologies), Part 3: Unified Speech and Audio Coding," 2012.