Audio Engineering Society
Convention Paper 5851
Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

Perceptual Optimization of the Frequency Selective Switch in Scalable Audio Coding

Miikka Vilermo[1], Sebastian Streich[1], Mauri Väänänen[1], Karsten Linzmeier[2], Bernhard Grill[2], and Ye Wang[1]

[1]1Nokia Research Center,
Speech and Audio Systems Laboratory,
33720 Tampere, Finland

[2]Fraunhofer-Gesellschaft IIS
91058 Erlangen, Germany

ABSTRACT

In a simple scalable audio coding scheme, there are usually two layers – a base layer and an enhancement layer. This paper presents a novel scheme with AMR-WB as base layer and AAC as enhancement layer. To optimally code the signal in the enhancement layer a frequency selective switch (FSS) control algorithm is described. The FSS determines whether the original signal or the residual of the original and base layer signals is sent to the enhancement layer in certain frequency bands. The proposed method introduces some advanced mechanism to the FSS and the quantization process as well as to minimizing the residual to achieve perceptually optimal result in the encoding process. These changes do not assume any modifications in the decoder.

1. INTRODUCTION

Scalability in audio coding is a property of the bit stream that makes it possible to create a meaningful representation of the original signal by decoding only a part of the bit stream. This is done often by defining a set of discrete scalability layers. The first layer is usually referred to as the base layer or the core codec and the additional layers as enhancement layers.

Scalable audio coding is an attractive scheme particularly for telecommunication applications. For example, a scalable bit stream is more resilient to different network capacities and traffic conditions because one can choose to only send the part of the bit stream that fits to the available bandwidth. Terminals can choose to decode the scalable bit stream according to their own capacities. The size of a scalable bit stream is usually smaller than that of non-scalable bit streams of the same bit rates. Also, bit rate scalability decreases the need for tandem coding. If combinations of speech and general audio codecs are used the scalable structure can even result in better quality, particularly with speech signals that

are difficult for general audio codecs [1]. Usually scalable audio coding schemes are not as efficient as their non-scalable counterparts at the same bit rates. This paper presents a novel scalable coding scheme that uses adaptive multi-rate wideband (AMR-WB) speech coder [2] as the base layer and MPEG-4 AAC [3] as the enhancement layer. In addition, some methods, which can be considered as perceptual optimization, are proposed to improve the efficiency of the scalable coding scheme.

2. AMR-WB

AMR-WB [2] is a speech codec that uses the code-excited linear predictive (CELP) coding model. It consists of nine coders each optimized for a different bit rate. The bit rates are 23.85 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.60 kbit/s. CELP coders work by sythesizing the speech. Excitation vectors are fed to the synthesis filter and the best vectors are chosen by using an analysis-by-synthesis search procedure in which the error between the original and synthesized speech is minimized according to a perceptually weighted distortion measure. The coder works internally with 12.8 kHz sampling rate and 20 ms speech frames. The bandwidth of the coder is 6.4 kHz except in the 23.85 kbit/s mode where the bandwidth is 7 kHz although the extra bandwidth is only filtered white noise. AMR-WB is an excellent speech coder and it has been selected for 3rd generation mobile communications. Thus, it will have a large installed hardware base.

3. AAC

AAC is a high quality modern audio coder. A block diagram of the encoder is presented in Fig. 1. The changes proposed in this paper are mostly related to the rate/distortion control process in the AAC encoder and therefore that part of the encoder is explained in detail. A good overview of AAC and the tools used in it can be found in [4].

The rate/distortion control process [3] tries to quantize the incoming signal so that the quantization noise always remains below the masking threshold while also making certain that the bit rate limits are not exceeded. This is done in the following manner (see Fig. 2-4). The transform domain coefficients are divided into scalefactor bands (SFB). Typically, there are 49 scalefactor bands. Each band contains a fixed number of coefficients. The lower frequency bands contain fewer coefficients (typically 4 coefficients per band) and the higher frequency bands contain more coefficients (up to 96). The quantization step size is the same for all coefficients within one scalefactor band. The division to scalefactor bands is related to the fact that the masking properties of the human ear are relatively constant inside these bands.

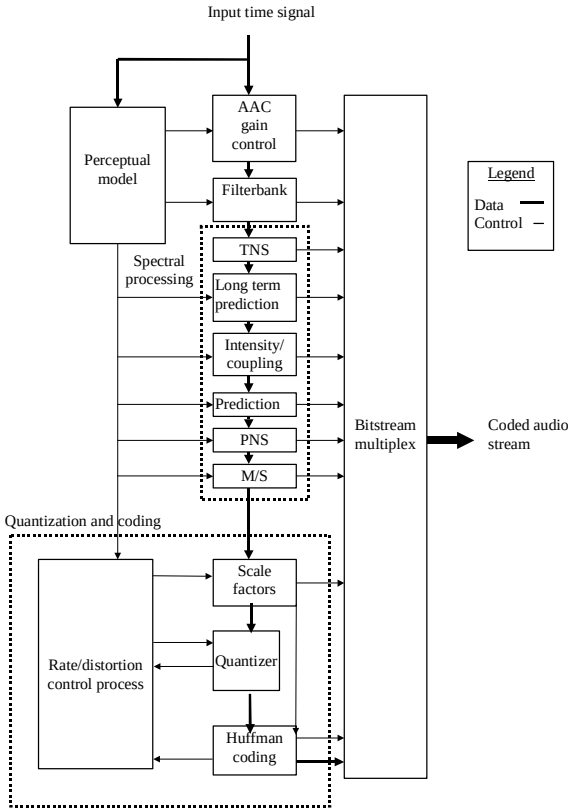

Figure 1. A block diagram of the MPEG-4 AAC encoder.

The equation used in the quantization is:

$$x\_quant = floor\left[\left(|x|2^{-\frac{1}{4}(scalefactor_i - common\_scalefactor)}\right)^{\frac{3}{4}} + MAGIC\_NUMBER\right] \quad (1)$$

where x is a coefficient in the transform domain, x_quant is the quantized coefficient, $scalefactor_i$ is a scaling factor that is the same inside a scalefactor band but can be different in different scalefactor bands. i refers to the scalefactor band that x belongs to. Typically $scalefactor_i$ can then have at most as many different values as the number of scalefactor bands e.g. i=0,…,48. These scalefactors define the distribution of quantization noise in the transform

domain. *common_scalefactor* is the same for all scalefactor bands in one frame. It controls the amount of bits needed in the frame. *MAGIC_NUMBER* is defined to be 0.4054 and it makes positive and negative numbers that have the same absolute value to be quantized to integers with the same absolute value.

The $scalefactor_i$ and the *common_scalefactor* can have a multitude of different values. Trying to find the optimal quantization by trying all the combinations for all scalefactor bands is computationally too sumptuous. Therefore AAC encoders usually use iteration loops within the rate/distortion control process. These loops try to iteratively find the optimum values for the $scalefactor_i$ and the *common_scalefactor* so that the quantization noise would be perceptually as small as possible.

The iteration process (Fig. 2) begins by calculating the amount of available bits in the current frame. All iteration variables are reset, i.e. $scalefactor_i$ and *common_scalefactor* are set to their initial values. A value called *quantizer_change* is also initialized. This value controls the way the *common_scalefactor* is changed. By changing it in bigger steps in the beginning, the process tries to arrive at the optimal result faster. Next, the process checks whether the incoming signal is all zeros. If so there is no need to quantize anything. Otherwise, the process continues to the *outer iteration loop* (Fig. 3).

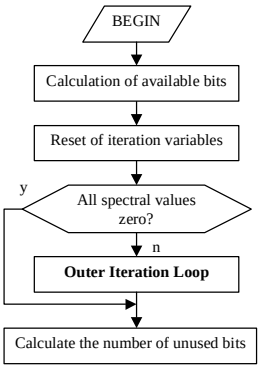

Figure 2. AAC iteration loop

The outer iteration loop first runs the *inner iteration loop* (Fig. 4). The inner iteration loop quantizes the signal with the current values of *common_scalefactor* and $scalefactor_i$. Then it counts the amount of bits needed to encode the quantized signal. When the inner iteration loop is called for the first time *quantizer_change* is set to 64. In subsequent calls of the inner iteration loop the *quantizer_change* is set to 2. The *quantizer_change* is added to the *common_scalefactor* and the inner iteration loop starts from the beginning. In subsequent loops, the *quantizer_change* is always halved. Every time the *quantizer_change* is added to the *common_scalefactor*, the quantization step size is also increased. This leads to a coarser quantization that takes fewer bits. When the amount of bits needed is less than the amount of bits available the *quantizer_change* is set to zero and the inner loop returns to the outer loop.

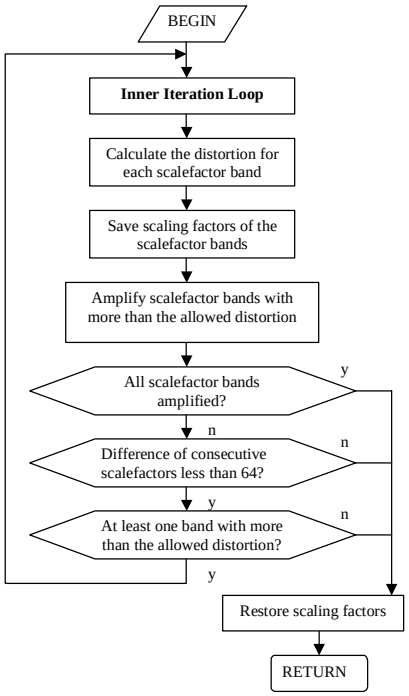

Figure 3. AAC outer iteration loop.

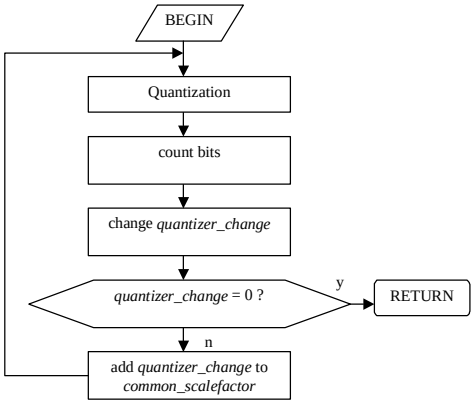

Figure 4. AAC inner iteration loop.

The outer iteration loop calculates the quantization error i.e. the distortion in each scalefactor band. Old scalefactor values are saved because the outer iteration loop may change the values so that they are no longer valid. In such event, the loop is terminated and the stored values are returned. If in some of the bands the distortion doesn't fall below the masking threshold, the $scalefactor_i$ for these bands i is increased. Increasing the $scalefactor_i$ decreases the quantization step size thus decreasing the distortion. If there still are bands with too much distortion, the process continues.

The process keeps on running so that the inner iteration loop makes the quantization step size bigger in all scalefactor bands by increasing the *common_scalefactor*. The outer iteration loop on the other hand tries to reduce the quantization step size in those bands where the distortion is greatest by increasing the $scalefactor_i$. The process tries to arrive in a situation where the quantization noise is perceptually evenly distributed across the entire frequency band. The process ends usually in the outer iteration loop when the quantization noise is below the masking threshold in all bands i.e. there is no band left with more than the allowed distortion. However, this cannot always be achieved. The process also ends if all $scalefactor_i$ have been amplified or if the difference between consecutive $scalefactor_i$ is greater than 64 because such scalefactors can no longer be written into the bit stream.

This is only one example of a possible implementation for the rate/distortion control process [3]. Nevertheless, this serves as a basis for the proposed improvements.

4. SCALABLE CODER

A typical scalable encoder structure is presented in Fig. 5, which is a two layer scalable codec for mono signals. The incoming signal going to the core codec is usually downsampled, especially if the core codec is a speech coder that tends to work at low sampling rates. After the decoded signal has been upsampled back to the original sampling frequency it is passed to the same filterbank used by the enhancement layer. The original signal is also passed through this filterbank. Two transform domain signals are formed: the original signal and the residual between the original and the core codec output. The residual signal is in effect the error produced by the core codec in the transform domain. Originally, this residual was calculated in the time domain but frequency domain systems work better [1].

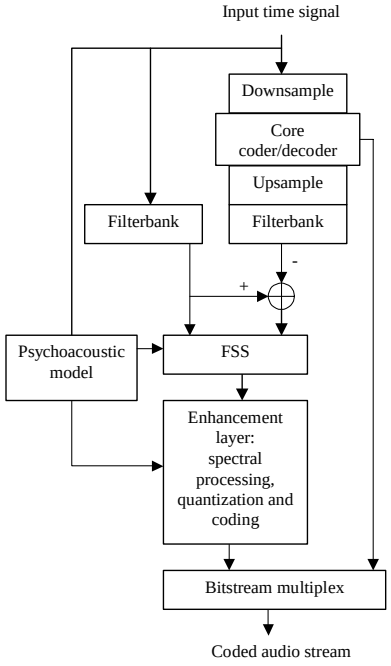

Figure 5. A block diagram of a two layer scalable encoder for mono signals.

At this point the scalable coder makes a critical decision: whether to try to improve the quality achieved by the core codec by encoding the residual or to scrap the core codec result completely and use the original signal instead, thus wasting all the advantage that might have been achieved using the core codec. This decision has to be done because sometimes the core codec achieves very poor quality due to the bit rate requirements and the residual would require far more bits to be coded than the original signal. The performance of the core codec may change from one frequency band to another and therefore the selection between the residual and the original is done in frequency domain bands. If AAC is used as an enhancement layer, it is natural to do this selection in scalefactor bands. This selection is done by the frequency selective switch (FSS) block.

Different implementations for the FSS have been proposed. Choosing the band with less energy is the obvious one, but better results have been achieved with the use of *perceptual entropy* [1]. Perceptual entropy is defined as [5]:

$$PE_i = \sum_{j=j_{low(i)}}^{j_{high(i)}} N(j) \quad (2)$$

where

$$N(j) = \begin{cases} 0 & , x(j) = 0 \\ \log_2\left(2\left|\text{nint}\left(\frac{x(j)}{\delta}\right)\right| + 1\right) & , x(j) \neq 0 \end{cases}$$

and $x(j)$ is a transform domain coefficient, j is an index telling which coefficient x is i.e. j tells the frequency of x. δ is the quantization step size that results in inaudible distortion, nint is a function returning the nearest integer, $j_{low(i)}$ and $j_{high(i)}$ are the upper and lower index of the scalefactor band i and PE_i is the perceptual entropy of scalefactor band i. The FSS calculates the perceptual entropy for each scalefactor band for both the residual and the original signal. Which ever has the smaller perceptual energy is selected for coding. The information about which option was selected is kept as additional side information. Based on this information the decoder then chooses in each scalefactor band whether to add the core layer to the enhancement layer when the enhancement layer signal is being decoded. It is worth mentioning that since the masking threshold only depends on the original signal it is the same for both signals, therefore δ is also the same for both signals.

The rest of the coding process works very much like with the non-scalable coder. In the rate/distortion control loop the masking threshold calculation must come from the original signal, otherwise the iteration process stays the same.

5. AMR-WB + AAC

MPEG-4 already has a scheme to combine a speech codec as the core codec and one or more layers of AAC [3]. That scheme uses a CELP coder that is part of the MPEG-4 standard. The CELP coder uses 8 kHz sampling rate and 30 ms long frames. To harmonize the combination, the AAC encoder has been modified in a few important ways. Since the coding is done on a frame-by-frame basis, the frame length of the AAC must match the frame length of the core codec. For this purpose, an alternative AAC frame length of 960 samples instead of 1024 samples is available. For example at 24 kHz sampling rate the 960 samples long AAC frame has got a duration of 40 ms. A super frame of 120 ms is created. It is made of three AAC frames and four CELP frames.

Incorporating AMR-WB to this scheme is a fairly straightforward task. AMR-WB has a frame length of 20 ms and it has externally 16 kHz sampling rate. If 24 kHz sampling is used in the enhancement layer a super frame of 40 ms is needed. It is made of one AAC frame and two AMR-WB frames. Other sampling rates can be used in a similar fashion.

6. LISTENING TEST

The AMR-WB + AAC scalable coder was implemented and tested in a listening test. The chosen test method was “multi stimulus test with hidden reference and anchors” (MUSHRA). The method makes possible to compare several different codecs together and gives good estimates of both relative and absolute performance levels. MUSHRA is a standard of European Broadcasting Union (EBU) [6] and International Telecommunications Union Radiocommunication Sector (ITU-R) recommendation BS.1534. The test was performed in a soundproof listening room with STAX electrostatic headphones. The test was performed by eight experienced listeners. The samples in the test are listed in Table 1. The samples are known to be very demanding and stressful for many aspects of audio coding.

Test Item	Description	Duration (seconds)
Es01	Vocal (Suzan Vega)	10
Es02	German speech	8
Es03	English speech	7
Si01	Harpsichord	7
Si02	Castanets	7
Si03	Pitch pipe	27
Sm01	Bagpipes	11
Sm02	Glockenspiel	10
Sm03	Plucked strings	13
Sc01	Trumpet solo and orchestra	10
Sc02	Orchestra piece	12
Sc03	Contemporary pop music	11

Table 1. Test material in the listening test.

The test used several different codecs and scalable coding schemes. These are listed in Table 2.

Tested codecs
hidden reference 1 (full bandwidth)
hidden reference 2 (bandwidth limited to 7 kHz)
hidden reference 3 (bandwidth limited to 3.5 kHz)
AAC (32 kbit/s, 10 kHz bandwidth, mono)
AAC (24 kbit/s, 7.6 kHz bandwidth, mono)
combination of AMR-WB and AAC (12,65 kbit/s AMR-WB + 19,35 kbit/s AAC, 10 kHz bandwidth, mono)
combination of AMR-WB and AAC (12,65 kbit/s AMR-WB + 19,35 kbit/s AAC, 9 kHz bandwidth, mono)
AMR-WB (12,65 kbit/s, 7 kHz bandwidth, mono)

Table 2. Codecs in the test.

The results of the test are presented in Fig. 6. The results show that the performance of the scalable combination at 32 kbit/s is close to the performance of the pure AAC at 32 kbit/s and the scalable codec was always at least as good as the pure 24 kbit/s AAC.

7. FSS MODIFICATION

Although the basic AMR-WB + AAC scalable coding scheme works quite well, its performance can be further improved with the following modifications.

7.1. Speeding up the Quantization

The rate/distortion control process can be computationally very demanding. Therefore, even minor optimization can speed up the quantization. One possibility here is to compare the signal to the masking threshold in the FSS. All those signal components that fall below the masking threshold can be set to zero already here. This may zero entire scalefactor bands thus speeding up the quantization. The same approach can be utilized also in non-scalable coders.

7.2. Verifying that Enhancement Layers Improve Quality

The FSS can use different criteria in its selection process. A simple energy based criterion is to always choose the option that has less energy. If the FSS uses the simple energy criterion, for choosing which bands use the residual signal and which bands use the original signal in the enhancement layer, then the following property of the AAC quantizer guarantees that the enhancement layers don't decrease the quality.

If the original signal energy in a scalefactor band is less than the residual signal energy then the original signal is selected in this scalefactor band. In the AAC enhancement layer the biggest error that can take place when encoding the original is when the signal is quantized to zero. The quantization error in this case is then equal to the original signal energy in this scalefactor band. Nevertheless, this energy is less than the residual signal energy that equals the error made by the core codec. Thus, the quantization of the original signal in the enhancement layer can only improve the base layer quality.

In the opposite case, the original signal energy in a scalefactor band is greater than the residual signal energy. Then the residual signal is selected in the FSS for this scalefactor band. Again, the greatest quantization error that can be made by the AAC enhancement layer is when the residual signal is quantized to zero. However, this error equals the base layer error and therefore coding the residual signal in the enhancement layer can only improve the quality.

The problem with the energy criterion is that it doesn't always give a very reliable estimate of the amount of bits needed to code a particular scalefactor band. The perceptual entropy measure usually works better. Unfortunately with the perceptual entropy criterion, the enhancement layer may actually end up increasing the error. A simple example of this is explained as follows. The original signal in the current scalefactor band is [2 2 2 8] and the residual signal is [4 4 4 4]. Let's assume that the wanted quantization step size is 1. Then the perceptual entropies of these scalefactor bands are

$$PE_i^{Original} = \sum_{j=1}^{4} \log_2\left(2\left|\mathrm{nint}\left(\frac{x(j)}{1}\right)\right|+1\right) \quad (3)$$

where
$x(j)=[2,2,2,8]$

and

$$PE_i^{Residual} = \sum_{j=1}^{4} \log_2\left(2\left|\mathrm{nint}\left(\frac{y(j)}{1}\right)\right|+1\right) \quad (4)$$

where
$y(j)=[4,4,4,4]$

Now $PE_i^{Original} \approx 11.1$ and $PE_i^{Residual} \approx 12.7$ and therefore the original signal is chosen in the FSS for quantization. However, if there are very few bits available in the enhancement layer the signal might be quantized to zero and the quantization error energy is then $2^2+2^2+2^2+8^2=76>64=4^2+4^2+4^2+4^2$ that is the error of the base layer. This means that by zeroing this SFB in the enhancement layer and decoding the base layer alone would result in smaller distortion.

In the opposite case where *x(j)* is the residual signal and where *y(j)* is the original signal, the FSS will try to encode the residual despite the fact that it would be a better choice to silence this SFB.

These extreme cases rarely occur. In most cases, choosing the signal with the lowest perceptual entropy still gives the best results. Nevertheless, it is worthwhile to find ways to overcome this problem. In the quantization, it is always possible to revert to two solutions. Firstly, one can choose to zero the enhancement layer and use the base layer signal alone in the enhancement layer decoder. The decoder can be forced to do this by zeroing the values in the SFB in the enhancement layer encoder and setting in the side information the correct bit to indicate that the residual signal was used in this SFB. The decoder then decodes zeros from the enhancement layer and adds them to the base layer signal. The distortion produced by this approach is equal to the residual signal energy. Secondly, the enhancement layer decoder can be made to zero the SFB completely by zeroing the values in the SFB in the enhancement layer encoder and setting in the side information the correct bit to indicate that original signal was used in this SFB. The error produced by this approach is equal to the original signal energy.

One solution to achieve this is to slightly change the AAC rate/distortion control process by modifying the outer iteration loop (Fig. 3). When in this loop the distortion for each SFB is calculated, instead of directly calculating the quantization error, a minimum of three values is chosen: the original signal energy in this SFB, the residual signal energy in this SFB and the quantization error in this SFB. Otherwise, the loops run as described earlier. If the quantization error is still bigger than either of the signal energies when the loop ends, then the quantized values in this SFB are replaced by zeros. In addition, the corresponding bit in the side information is set to indicate that the residual signal was coded if the residual signal energy was the smallest, or the bit is set to indicate that the original signal was coded if the original signal energy was the smallest. This way the coder retains the property of never increasing the error in the enhancement layers.

Zeroing some of the SFBs in the end of the quantization loop in the enhancement layer frees some extra bits. These bits can be used in following frames by saving them in the bit buffer or they can be redistributed in the current frame. This can be done by running the quantization loop a few extra iterations starting from the current values and not making the comparison between the signal and the quantization error energies in the extra iterations.

7.3. Speech Codecs Code Speech Well

Speech codecs are strange in the sense that they achieve perceptually very good quality although the coded signal might be significantly different from the original signal. In scalable coding where the base layer is a speech codec, this may cause problems with speech signals when the residual signal is large even though the core codec's output sounds very similar to the original signal. Then the enhancement layer wastes bits trying to improve SFBs where the signal actually sounds good. In theory, the psychoacoustic model should be able to estimate the masking threshold so well that this weren't a problem but in practice the phase distortions, time alignment problems and other features of speech codecs make this a difficult task.

A simple solution to overcome this problem is to divide the frequency band into two halves. The core speech codec takes care of the lower half and the enhancement layer codec encodes the high frequencies. While this works well for speech signals, this approach has problems with other signals. Thus, a more adaptive system is needed.

A good estimate of whether the incoming signal is speech or not can be achieved by calculating the total error over the masking threshold made by a speech codec over the bandwidth of the codec. If the error is small, the signal is most likely speech or something that the codec encodes well.

In scalable coding where the base layer is a speech codec, an estimate of the speechlike quality of the input signal can be used to modify the quantization

process in the enhancement layer. When in the previous proposed change the quantization error was compared directly to the original signal energy and the residual signal energy this time the comparison is done between the signal energies and a weighed quantization error. If the input signal is speechlike the quantization error is weighed with a value > 1. If the signal is not speechlike, the comparison remains unchanged. This way the enhancement layer is forced to focus on the SFBs having very large distortions when the incoming signal is speech. This way bits can be saved in SFBs where the speech core codec probably achieves perceptually good quality although the distortion were a little above the masking threshold. It is clear that this method is not restricted to speech core coders. It works with all types of core codecs.

A block diagram of the proposed changes is presented in Fig. 8 and Fig. 9. In Fig. 8 the quality of the core codec signal is estimated by summing the core codec error in one frame. If the error made by the core layer is below the masking threshold in a SFB then that SFB is zeroed. In Fig. 9 the quantization error is weighed by the quality factor. The quality factor is > 1 if the quality achieved by the core layer is high. This error is compared to the original signal energy and the residual signal energy. The minimum of these is used to estimate the goodness of the current quantization. If quantization error is still greater, after the quantization loop has finished, then the SFB is zeroed and the FSS side information is modified.

7.4. Quality improvement by predistortion

The perceptual quality of the output of an audio encoder/decoder system using the kind of large step scalability described above strongly depends on how much bit rate can be saved compared to the simulcast mode by exploiting similarities in the spectra of core and enhancement layers. The expression “simulcast mode” means that the enhancement layers are encoded independently of the core layer. In this case, related to the bit rate of a non-scalable encoder/decoder, the entire bit rate of the scalable bit stream will be increased by the bit rate of the core layer, assumed that the quality of the audio output shall be maintained. On the other hand, if a maximum of similarity in the spectra can be exploited, there won't be an increase of bit rate at all.

This consideration shows that it is highly desirable to have as much analogy in the spectra of core and enhancement layers as possible. At this point two details of the scalable AAC coder with AMR core come into play: These are the application of pre- and postfiltering in the AMR core coder on the one hand and the use of an MDCT as a transformation to calculate the spectra of time signals for further processing in the scalable AAC coder on the other hand.

Pre- and postfilters are used in most time domain coders. One application scenario is e.g. to filter out the constant component and/or high frequency content of a signal. Whereas such kind of filtering won't change the energy content of the signal's amplitude spectrum in the frequency range of interest, it will in general introduce phase distortion. This kind of distortion is scarcely audible for the human ear, but it changes the shape of an MDCT spectrum calculated out of the so treated signal. This leads to an increase in the differences of the spectrum of the unprocessed signal representing the input of the AAC part of our scalable coder and the spectrum of the signal which passes the AMR encoding and decoding in the core coder, and therefore, taking into consideration what has been stated above, will deteriorate the performance of the scalable encoder/decoder system.

In order to avoid the prescribed problem, two kinds of solutions can be thought of: Either to undo the phase distortion introduced by the core coder by postfiltering the output signal of the AMR coder using a filter with a complementary phase response or to predistort the simulcast input signal of the AAC enhancement layer using a filter with a phase response similar to that of the core coder's phase distortion. As in either case, only the phase of the particular signal shall be altered in order to keep the changes inaudible, the best choice to do this is to use an allpass filter.

There have been two main reasons for not to choose the postfiltering of the core coder signal in our implementation: First of all a stable allpass filter causing a complementary phase response to that of the core coder could not easily be found. And second, it would have been necessary to do this kind of distortion in both the encoder as well as the decoder, in order to obtain the same core layer input signal for the frequency selective switch (FSS). Predistorting the simulcast AAC encoder input signal, this needs not necessarily to be done.

Figure 10 shows the block diagram of the scalable encoder including the predistortion tool.

8. TESTING THE IMPROVEMENTS

Improvements in the FSS were informally tested with a scalable codec structure with AMR-WB as the core layer and an AAC like coder as the enhancement layer. This test showed clear improvements over the system without them, but a more formal test with a state of the art AAC encoder like the one used in the AMR-WB + AAC test earlier in this paper would be needed to establish the importance of these changes. Some of the improved quality may be attributed to the low quality of the used enhancement layer codec and the ease of finding ways to improve it.

In Fig. 10 are depicted the results of a BS.1284 [7] conform listening test comparing the performance for the scalable AAC coder with AMR core with and without usage of the prescribed predistortion method. Positive values correspond to an improvement by the use of predistortion. The scalable configurations in test were "12.65 kbit/s AMR-WB core + 19.35 kbit/s mono AAC enhancement layer" and "15.85 kbit/s AMR-WB core + 16.15 kbit/s mono AAC enhancement layer".

The results of the listening test show that for most of the signals out of the standard MPEG test set the sound quality was clearly improved by the use of predistortion.

9. FUTURE WORK

To find out the significance of the proposed changes in the FSS they would need to be implemented in a state of the art scalable coder as was done in the predistortion test. The efficiency of scalable coders still lags behind non-scalable coders. Ways to utilize the side information of lower layers in higher layers might help to narrow the gap. Also a better understanding of the tolerances in the human auditory system would help in choosing what parts are important to encode in the enhancement layers.

10. CONCLUSIONS

A novel scalable coder scheme was presented. This scheme with AMR-WB as the core codec and AAC as the enhancement layer achieves good quality. This was verified in a listening test. Scalable codecs still lag behind non-scalable codecs. Four ideas were proposed to narrow this gap. The first of these ideas focused on making the quantization process faster. The second idea removed the possibility of degrading the audio quality with enhancement layers when using advanced frequency selective switches. The third idea makes better use of the knowledge that the core codec is a speech codec to improve the speech quality in the enhancement layers. The fourth idea focuses on maximizing the similarities between the base and enhancement layers.

11. REFERENCES

[1] Grill, B., "A Bit Rate Scalable Perceptual Coder for MPEG-4 Audio," presented at the AES 103rd Convention, New York, USA, 1997 September 26-29.

[2] 3rd Generation Partnership Project, "TS 26.190 V5.0.0, AMR Wideband Speech Codec," 2001.

[3] ISO/IEC 14496-3 International Standard, "Information technology — Coding of audio-visual objects — Part 3: Audio," 2001.

[4] Kunz, O., Brandenburg, K., "An Overview of MPEG-4 Audio," presented at the AES UK Conference Audio: The Second Century, UK, 1999 June.

[5] Johnston, J., "Estimation of Perceptual Entropy Using Noise Masking Criteria," presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-88), vol. 5, pp. 2524-2527, New York, USA, 1988 , April 11-14.

[6] European Broadcasting Union, "MUSHRA — EBU method for subjective listening tests of intermediate audio quality," Draft Technical Recommendation BMC 607 (Rev. 1) B/AIM 022 (Rev.9), Technical Department, 2000, January.

[7] ITU-R, "Methods for the Subjective Assessment of Sound Quality — General Requirements," International Telecommunications Union, Radiocommunication Assembly, Recommendation BS.1284, Switzerland, Geneva, 1998.

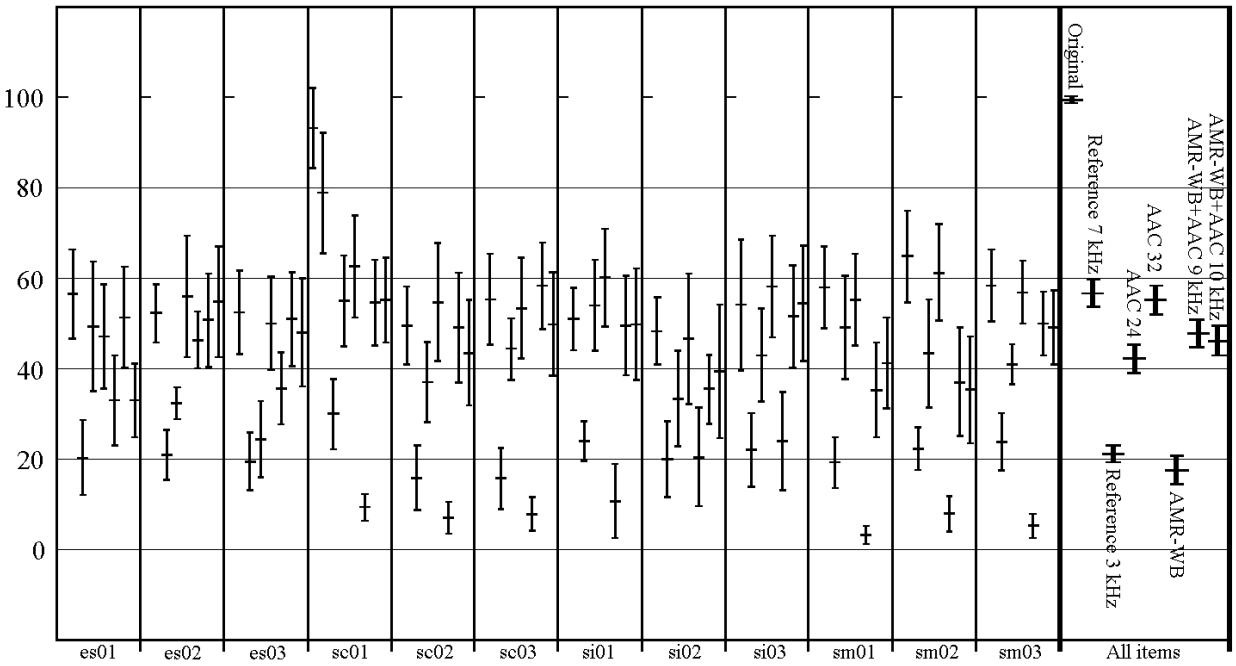

Figure 6: MUSHRA listening test comparing performance of different coding schemes with different samples. Codecs are in the same order in all columns.

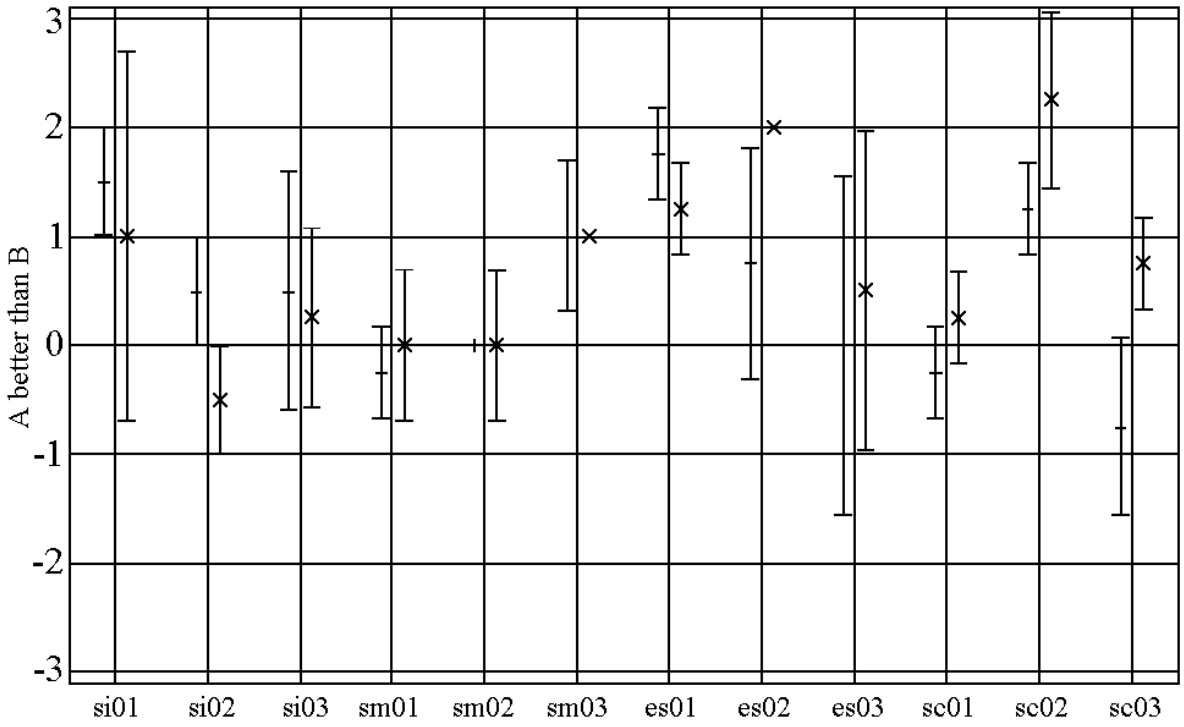

Figure 7: BS.1284 listening test comparing performance with and without use of predistortion. 12.65 kbit/s AMR-WB + 19.35 kbit/s AAC and 15.85 kbit/s AMR-WB + 16.15 kbit/s AAC (marked with x) were used in the test. Positive values indicate improvement with the predistortion method.

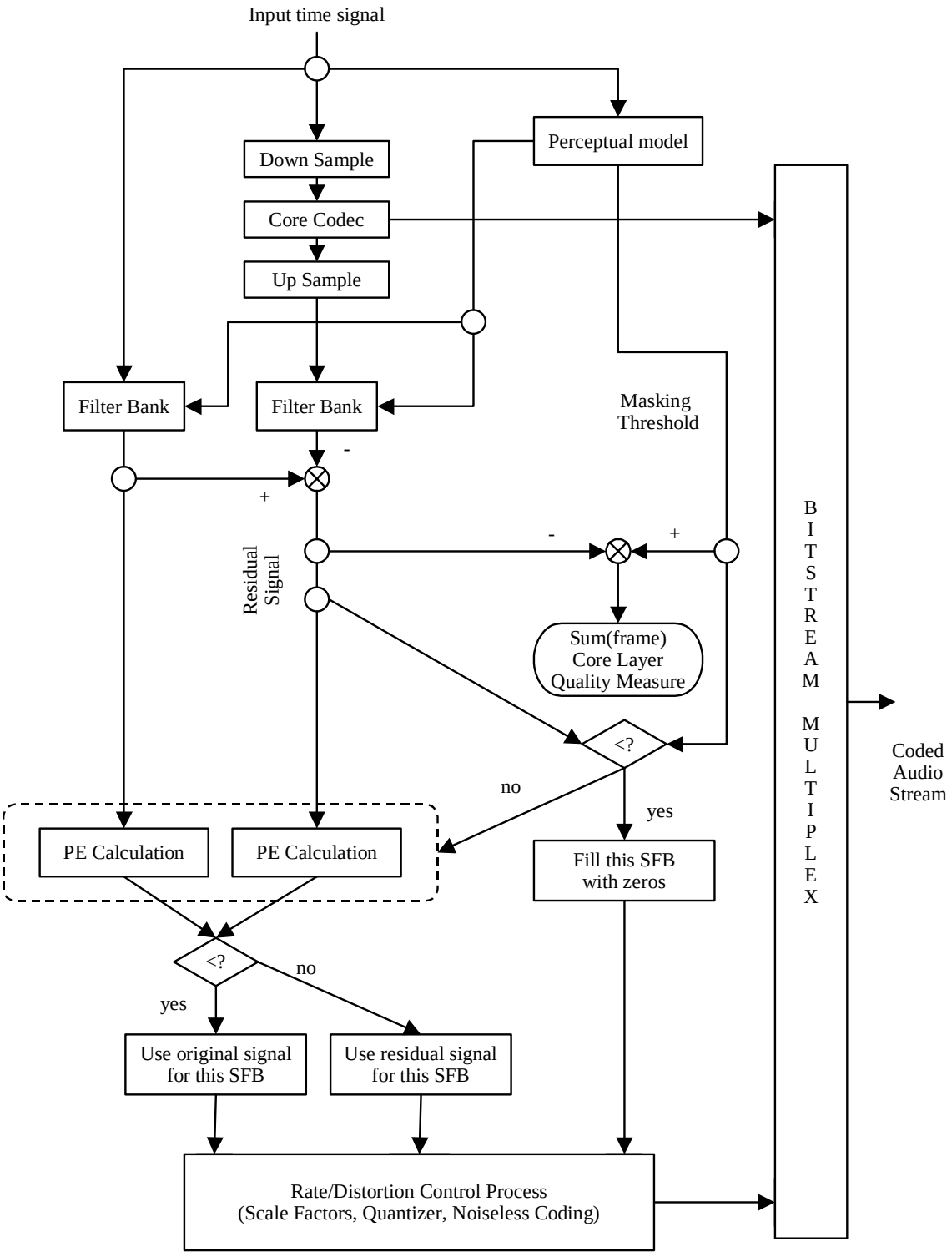

Figure 8. Block diagram of the scalable encoder with the proposed changes.

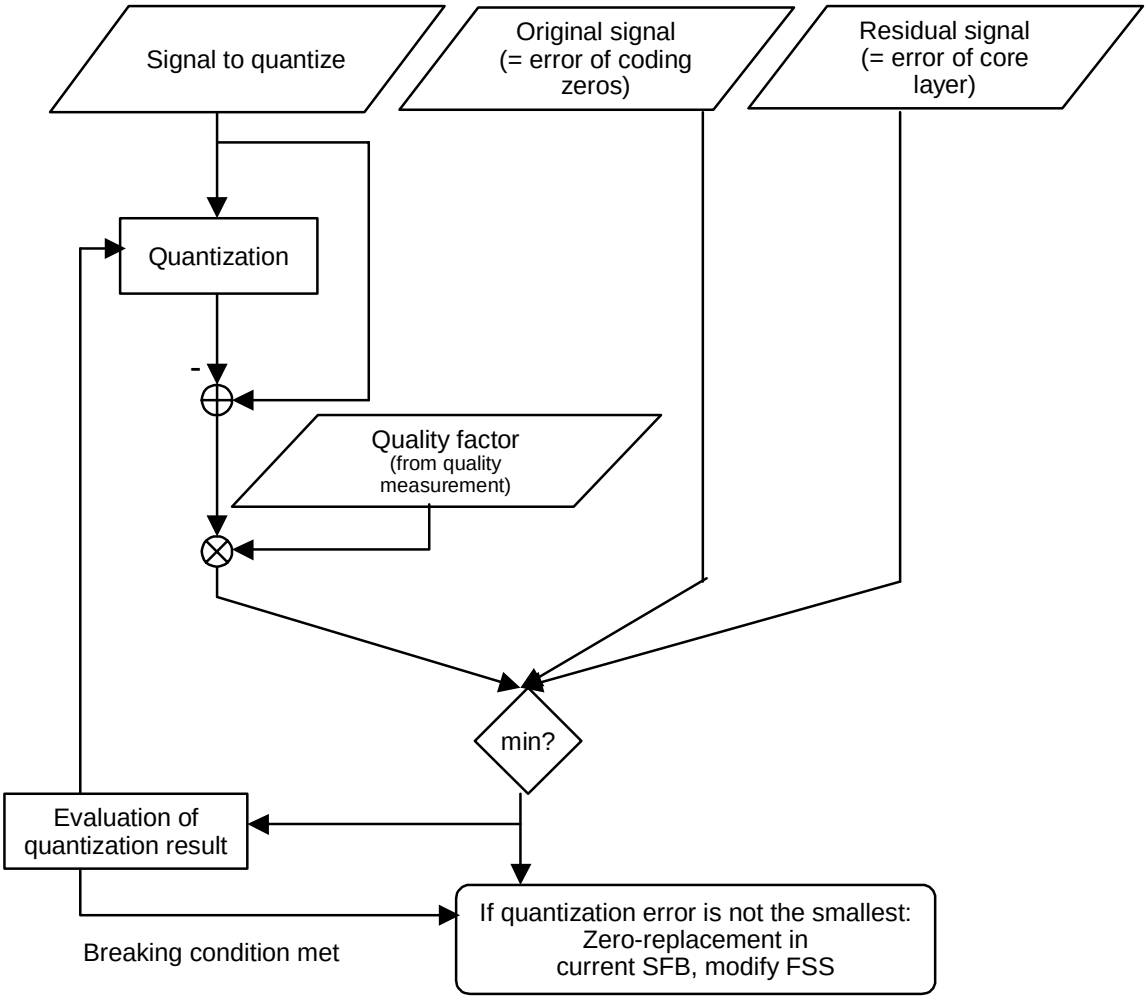

Figure 9. Block diagram of the scalable encoder quantization process with the proposed changes.

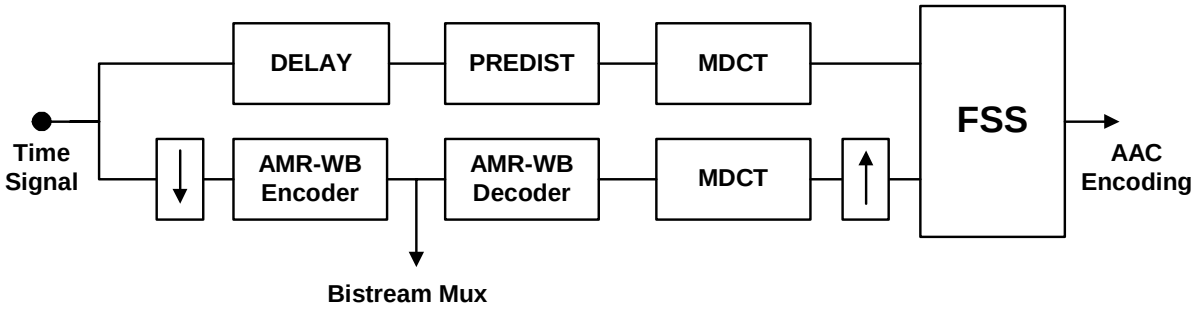

Figure 10: Block diagram of the scalable coder using predistortion.