A SCALABLE SPEECH AND AUDIO CODING SCHEME WITH CONTINUOUS BITRATE FLEXIBILITY

Balázs Kövesi, Dominique Massaloux and Aurélien Sollaud

France Telecom R&D, 2, Av Pierre Marzin, 22300 LANNION – France
{balazs.kovesi,dominique.massaloux,aurelien.sollaud}@rd.francetelecom.com

ABSTRACT

Networks are getting more and more heterogeneous. Scalable codecs are especially suited for such a context as they permit to lower the bitrate in a simple way, at any point of the transmission, for adaptation to network conditions and to the terminal capacities. Classically, scalable codecs are organised in layers and scalability is obtained by sending more or less layers to the decoder. The obtained granularity depends on the layers sizes, and the available bitrates are fixed and limited in number. This paper presents a novel scalable audio coding scheme where the bitrates vary continuously between a minimal and a maximal value, allowing free modification of the bitrate. With this novel approach, all bitrates are valid, sending even one more bit results in different output signal with statistically growing quality. Test results show that this method provides as good or even better quality than that of a not scalable version.

1. INTRODUCTION

The main topic of this paper is scalable coding, where the coded information is organised in layers: A core layer contains the minimum information needed to obtain a basic quality decoded signal. Each successive layer contains new information on the same signal part and improves the previously obtained quality. There are two main ways to improve the preceding layers quality: either by refining the quantization precision of already decoded parameters (residual coding) or by sending new parameters.

Quite commonly, the core layer coding scheme is different from that of the other layers: the core codec techniques provide their best performance at low bitrates, whereas extension layers use methods more suitable for higher bitrates. This allows covering a large range of bitrates and qualities, with sufficient quality at relatively low rates and high quality at high rates. A typical example is the CELP – transform coding pairing [1].

Scalable encoders process and send the full range of layers, but the decoders do not necessarily receive all the layers, since the last layers can be dropped at any point of the transmission or storage chain. The decoders process the received layers with an obtained quality roughly proportional to the number of decoded layers.

Thanks to this flexibility, scalable codecs are especially adapted to heterogeneous (in terms of access and terminal capacities) networks. This feature has been for instance exploited in audioconferencing services over IP involving various types of clients (over xDSL, dial-up connections, etc..). Another typical usage is database consulting, where the signal is stored in the highest quality mode and the sent bitrate is adapted to the link or to the client capacity. Congestion handling can also benefit of the bitstream scalability: when the channel bandwidth becomes insufficient the bitrate of some communications can be reduced in the network to reduce data loss.

Scalable codecs have been proposed for a long time in the art: the embedded schemes like ITU-T G.727 (16/24/32/40 kbit/s) for telephone bandwidth or ITU-T G.722 (48/56/64 kbit/s) for wideband speech are examples of scalable coders. For such coders, all the layers produce signals of the same bandwidth that limits the quality. And the scalability is obtained at the price of degradation in terms of performances when compared to fixed rates schemes.

More recently, the MPEG4 Audio standard also proposed an embedded version of the MPEG CELP coder [2], where up to 3 enhancement layers can be used (2 kbit/s in the narrowband version and 4 kbit/s in wideband). Test results comparing the basic version (not scalable) of the MPEG CELP codec with a scalable version (core layer of 6 kbit/s), at 8 and 12 kbit /s have shown that both at 8 and 12 kbit/s, the scalable version has a quality significantly lower than the basic version at the same bitrate.

Many other scalable codecs increase the bandwidth with added layers. For instance, the MPEG CELP codec has also a bandwidth scalable version [2]. Another bandwidth scalable standard is the MPEG AAC-BSAC [3], with quite fine bitrate graduation of 1 kbit/s.

The MPEG4 CELP-AAC (ISO 14496-3) standard is a good example of a scalable structure with two different techniques in the core and extension layers.

In all those existing scalable codecs, the bitrates allocated to the layers are defined in advance, and the granularity depends on this bitrates. Test results also show that the performance of such scalable codecs is generally lower than that using the same coding technique optimised for the same fixed bitrate. There are several reasons (depending on the codec):

- parameters quantization is split between the layers which is less efficient than a unique quantization,
- enhancement layers include side information to describe their content, which increases the bitrate without improving the quality,
- prediction and backward adaptation are performed using the core decoded parameters to keep the encoder and decoder synchronized when the decoder only receives the lower bitrate.

Generally, the finer is the granularity, the higher this loss is. That is why the number of available rates is often limited. The present paper describes a new technique that does not suffer from the abovementioned limitations and shows good performance compared to equivalent fixed rate schemes, with very fine grain flexibility.

The rest of the article is organised as follows: the principle of the new coding scheme is first explained. Two examples of scalable codecs according to this method are then presented. Prior to this, a short description of the transform codec that is at the basis of the two scalable codecs is given. Finally some quality evaluation results are given.

2. A NEW SCALABLE SCHEME

In this new scalable coding scheme the encoding is performed for the highest bitrate. The bits of the obtained coded parameters are divided in two groups. The first group, put at the beginning of the bitstream, contains the coded parameters that are absolutely necessary to obtain a minimal and acceptable quality. The rest of the bits are ordered and put in the bitstream according to their importance order. This order has to be known also by the decoder. The core layer consists of first group bits and possibly some of the most important bits of the second group. When fixed length quantization is used the minimal bitrate estimation is straightforward. When variable rate entropic quantization is used, the minimal bitrate is chosen as the statistically lowest bitrate where even in the worst case enough parameters are received to produce the target minimal quality. In this case the core layer contains a variable number of encoded parameters, and on average more than the strictly necessary number to achieve the minimal quality.

The encoder sends the whole or part of the bitstream, which can be further shortened during the transmission. The decoder decodes the received parameters, and regenerates the missing ones from the received ones. Of course if more parameters are available, the decoded quality is better, since fewer parameters need to be estimated.

When the bitstream truncation leaves the last parameter index incomplete, either this incomplete index is ignored or, if the quantization method allows it, the corresponding parameter is partially decoded.

Note that in the case of entropy coding, for a given bitrate, the number of the decoded parameters changes at each frame. For a long sequence, by sending one more bit per frame, the number of entirely decoded parameters statistically increases and the quality increases accordingly.

The two examples detailed further are based on a transform codec called in the following the "TDAC codec" (from "Time Domain Aliasing Cancellation") [4]. The next paragraph gives a brief overview of this technique.

3. THE TDAC CODEC

The TDAC codec computes the MDCT transform of the input frame. The obtained spectral coefficients are grouped in frequency bands of unequal width, with larger bands at higher frequencies. A perceptual module computes a masking curve, sets to 0 the masked coefficients and identifies (if any) the fully masked bands. For each non-masked band, a scale factor is computed, quantized and Huffman coded. Masked bands are signalled by a dedicated Huffman code. Based on the de-quantized scale factors the perceptual module estimates the perceptual importance of the bands. The bit allocation module uses this information to distribute the remaining bits among the unmasked bands. The normalised fine spectral structure of the bands is finally quantized using a very flexible vector quantizer.

This MDCT transform coding scheme is particularly well suited to the above presented new scalable coding scheme, especially for the following reasons:

- The largest part of the bitstream is used to transmit the fine structure of the spectrum. These parameters can be ordered according to their perceptual importance, this information being available in the basic version of the codec as mentioned above.
- The perceptual importance is computed from the de-quantized scale factors, available both at the encoder and at the decoder sides. So no extra bit is needed to transmit this information.
- The fine spectral structure of the less important bands can be easily and efficiently regenerated in function of the received neighbouring bands, with a good perceptual quality.

4. A SCALABLE WIDEBAND TDAC CODEC

The scalable codec presented in this paragraph is based on a 10 ms frame length version of the TDAC codec, for 16 kHz sampling frequency signals. This short frame length

has been chosen to reduce pre-echoes at low bitrates and to maintain the algorithmic delay low and improve interactivity for conversational applications. Larger delays versions have also been developed, based on this first short frame version.
Computation of the MDCT transform coefficients, scale factors quantization and Huffman coding is performed as explained in section 3. The partial bitrate corresponding to the spectral envelope varies between 5.5 and 12 kbit/s, those bits form the first group mentioned in section 3.
The remaining bits are distributed among the unmasked bands according to their perceptual importance, for a total bitrate of 32 kbit/s.
Corresponding to the bitrate of the first group, the minimal bitrate has been chosen equal to 12 kbit/s, which constitutes the core layer. In this way, all scale factors are transmitted even in the worst and very low probability cases. In fact, this core layer generally also contains the first most important bands at least.
In normal operating mode the decoder receives between 120 and 320 bits per frame. After decoding the scale factors the same bit allocation as done at the encoder is computed, and the rest of the received bits corresponding to the normalized spectrum of the frequency bands can be decoded. For the frequency bands above 1000 Hz, when the normalized spectrum information is not received, it is regenerated: for voiced frames by mirroring the normalized spectrum of the previous bands, for unvoiced frames by generating a random noise (see figure 1). For voiced frames special care has been taken to avoid breaking the harmonic structure.
It has been found that below 1000 Hz , the regeneration process creates a high risk to introduce annoying artefacts, that is why it is preferred to set to zero the non received coefficients in this frequency region. In fact, these bands are either masked or judged important by the perceptual unit and in this case they are among the first sent bands.

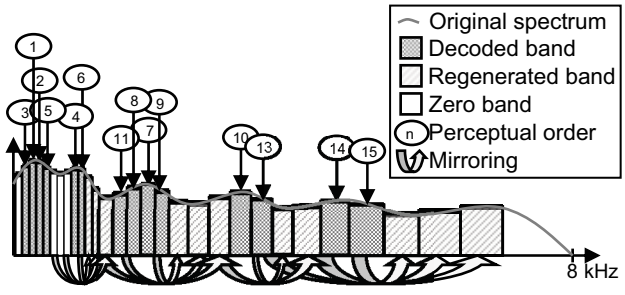

Figure 1: Intermediate bitrate decoding of a voiced frame

The regeneration of the non received spectrum parts is very important. Without it, at lower bitrates, holes appear in the spectrum. Because of the variable importance order and the variable rate Huffman coding, their number, sizes and positions are different from frame to frame, resulting in a very annoying noise. This artefact can be drastically reduced by the applied parameter regeneration.
Note that when less than 120 bits are received by the decoder; it is still probable that the frame can be decoded correctly, depending on the performance of the Huffman coding. Even if only a part of the scale factors is received this can be exploited by the frame erasure concealment procedure to correct its model.

5. A SCALABLE CELP-TDAC CODEC

In this second example, the ITU-T G723.1 codec at 6.4 kbit/s (including VAD and bitrate information) is used as core codec. The frame length of this narrow band, low bitrate CELP speech codec is 30 ms. The enhancement layer for two consecutive G723.1 frames is a wideband, 60 ms version of the TDAC codec, containing three 20 ms MDCT analyses. This layer operates in the range 0-25.6 kbit/s, resulting in a global bitrate range of 6.4 – 32 kbit/s.
The analyses are made on the residual signal obtained by subtracting the upsampled core layer decoded signal from the original signal. The lower frequency bands (below 3450 Hz) contain the spectral correction of the core codec output signal while the higher bands contain the input signal.
The enhancement layer bits can be divided into three groups and are put in the bitstream in the following order:

A. The scale factors of the higher frequency bands (> 3450 Hz), in increasing order.
B. The scale factors of the lower frequency bands, in increasing order.
C. The fine spectral structure of all frequency bands, ordered according to their perceptual importance.

Corresponding to this bitstream organisation, the decoder processing depends on the number of the received bits:

1 If the core layer is not entirely received the decoder toggles to the frame erasure concealment mode.
2 If only the core layer is received, the output of the G723.1 codec is produced (cf. figure 2). The resulting minimal quality is correct for speech samples but not adapted to music signals.

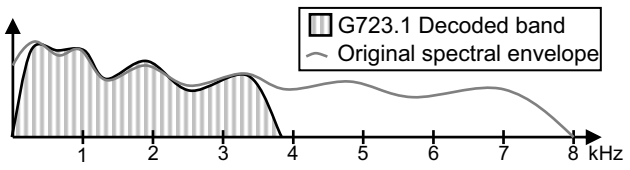

Figure 2: Spectrum of the core layer output

3 If a part of the group A bits is also received, some of the higher bands scale factors can be decoded which allows part of the higher band regeneration by

mirroring the fine spectral structure of the known lower bands (figure 3). The quality is better for speech samples than in case 2, due to the larger band, but still limited for music.

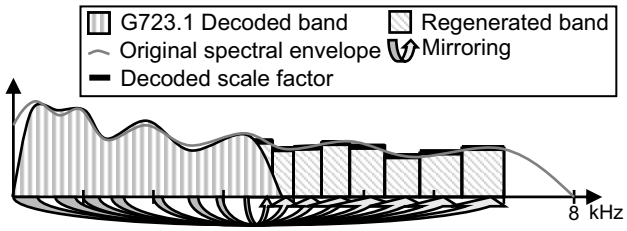

Figure 3: Regeneration knowing scale factors only

4 Receiving the group B bits without knowing the fine spectrum structure of these bands is hardly exploited. In this case this part of the received bitstream is in fact unused. However, in the bitstream, from frame to frame, this group of bits has different size and position, (due to Huffman coding of the scale factors) so on average all bitrates are useful.

5 The decoder can also decode some fine spectral structure information. In the lower bands it is used to correct the spectrum of the core codec. This is particularly important for music samples. In the higher bands the decoded coefficients replace the regenerated ones and allow better estimations for the neighbouring, not decoded spectrum parts (see figure 4). The quality increases continuously with the bitrate for all kind of signals.

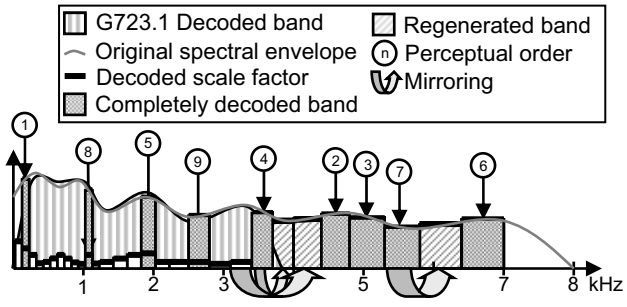

Figure 4: Intermediate bitrate decoding of a voiced frame

6 TEST RESULTS

A formal test involving 24 trained subjects was performed to evaluate the quality of the scalable codec presented in paragraph 5. The 24 and 32 kbit/s modes were compared to the basic TDAC codec and to the ITU-T G722.1 standard at the same bitrates. The tested scalable codec was always at least equivalent to these reference codecs and was even significantly better than the basic TDAC for noisy speech samples at 24 kbit/s.

Informal test results also confirm that at lower bitrates the presented scalable coding scheme performs equally or even better than the equivalent non scalable version where the bit allocation is made for the given bitrate.

7 APPLICATION

This codec is used in an audio-conferencing solution for IP networks developed by France Telecom R&D. This solution is based on a client – server architecture. The server simply replicates the coded audio streams sent by the software clients.

Thanks to the scalable codec, non homogeneous conferencing sessions are possible. The server truncates the bitstreams in such a way that the best possible quality is provided. The bitrate management is performed according to rules that take into account the access bitrate and the number of attendees in the session.

This application has been integrated into a complete service that associates Presence Management and audio conferencing. This service is currently deployed and accessible by any Internet Access (see [5]).

8 CONCLUSION

A very flexible and efficient scalable coding scheme was presented. With this method, scalability is no more a handicap in quality point of view. These new codecs have the same performance as the non-scalable ones at high bitrate and can even outperform them at lower bitrates. Therefore they can also be successfully employed in a non-scalable, fixed bitrate, mono-rate or switchable multi-rate context.

Having a limited number of bits, it seems to be more efficient to encode well the most important parameters and regenerate the remaining ones than encoding all parameters with higher distortion for each. This means that, for a fixed bitrate, the best result is obtained by doing bit allocation according to a higher bitrate.

9 REFERENCES

[1] SA. Ramprashad, " Embedded coding using a mixed speech and audio coding paradigm," International Journal of Speech Technology, Vol.2, May 1999, PP.359-372.

[2] B. Edler, " Speech Coding in MPEG-4," International Journal of Speech Technology, Vol.2, May 1999, PP.289-303.

[3] SW. Kim, SH. Park, YB Kim, " Fine grain scalability in MPEG-4 Audio," AES Conv. Paper 5491, Sept.2001.

[4] Y. Mahieux, J.P. Petit, " High quality audio transform coding At 64 kbit/s," IEEE Trans. on Com., Vol.42-11, Nov.1994, PP.3010-3019.

[5] URL **http://wanadooaudio.wanadoo.fr/**