DYNAMIC STRATEGY FOR WINDOW SPLITTING, PARAMETERS ESTIMATION AND INTERPOLATION IN SPATIAL PARAMETRIC AUDIO CODERS

Julien Capobianco[(a)], *Gregory Pallone*[(a)] *and Laurent Daudet*[(b)1]

[(a)]France Telecom Orange Labs/TECH/OPERA, Av. Pierre Marzin, 22307 Lannion Cedex, France
[(b)]Institut Langevin and Paris Diderot University, ESPCI 10 rue Vauquelin 75005 Paris, France

ABSTRACT

In most parametric stereo audio coders, sets of spatial parameters are extracted from the audio channels in a time-frequency domain. In order to reduce the amount of data, the parameters plane is highly down-sampled, and transmitted together with a mono downmix. Then, in the decoding process, it is necessary to interpolate the upmix matrix computed from these parameters. Usually, this is done in the same way for each portion of signal, regardless of its nature. In this article, we propose a dynamic strategy of window splitting, estimation of the parameters and interpolation of the upmix matrix based on transient detection in the audio signal. Subjective tests show an improvement when applied to the new stereo parametric tool from MPEG USAC.

Index Terms— Parametric audio coding, stereo

1. INTRODUCTION

Recently, interest for multichannel audio has grown as 3D video and immersive multimedia have spread among a large variety of devices, including devices with limited bandwidth and storage like mobile devices. For handheld devices, low bitrate audio is required especially as immersive content is extremely bandwidth consuming.

In the last decade, low bitrate multichannel audio coding has clearly converged to spatial parametric techniques. It began with Intensity Stereo [1], followed by more sophisticated models like Binaural Cue Coding (BCC)[2], Parametric Stereo (PS) [3][6] and MPEG Surround (MPS) [4]. USAC [5], the latest MPEG audio coder, embeds a derived version of MPEG Surround for the stereo case by integrating and updating features from PS. However, these approaches do not reach transparency on most audio material and introduce several artifacts on critical samples.

This article shows that perceived audio quality can be improved by a dynamically-adapted window splitting, and corresponding parameters estimation and interpolation, based on detection of relevant transients inside the audio signal.

Transients have an important role in localization, notably by the precedence effect [10], and are often associated to a sudden change in a spatial scene. Subjective tests we conducted on PS put in light spreading and bad localization of transients. PS or USAC, by some aspects, could allow implementing the dynamic strategy we propose at the encoder, but the decoder part prevents a dynamic strategy based on signal cues like the transient to be used.

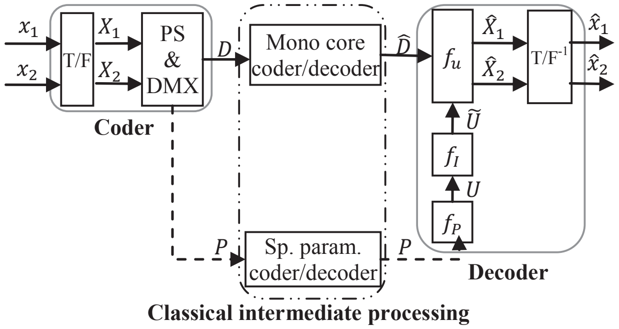

Figure 1. Block diagram of a general spatial parametric coder

The paper outline is as follows. Firstly, common structure of spatial parametric coder is presented in section 2. Then, our technique is explained in Section 3. In section 4, experimental setup and subjective test results are presented and we discuss about performance aspects of our technique. Finally, a conclusion is drawn in Section 5.

2. PARAMETRIC AUDIO CODING BASED ON TIME-FREQUENCY ANALYSIS

Most spatial parametric models work in a time-frequency domain, reminiscent to the human auditory system analyses an audio scene. In this part, we introduce general structure and notions for such coders.

A general structure of typical stereo parametric coders is shown on Figure 1. The two input stereo signals x_1 and x_2 of the time domain input stereo signal are first transformed into time-frequency signals X_1 and X_2. Then, a number of spatial parameters P are estimated and a mono signal D (called downmix) is created from X_1 and X_2. At the output of the spatial parametric coder, there is only the bitstream corresponding to the mono audio signal plus a stream of spatial parameters (with a bitrate substantially lower than the one of the mono audio signal). For the sake of clarity, we here do not consider the mono core coder used to encode the downmix, and the associated transformations which may appear before feeding the mono coder.

[1] LD is on a joint affiliation with the Institut Universitaire de France

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

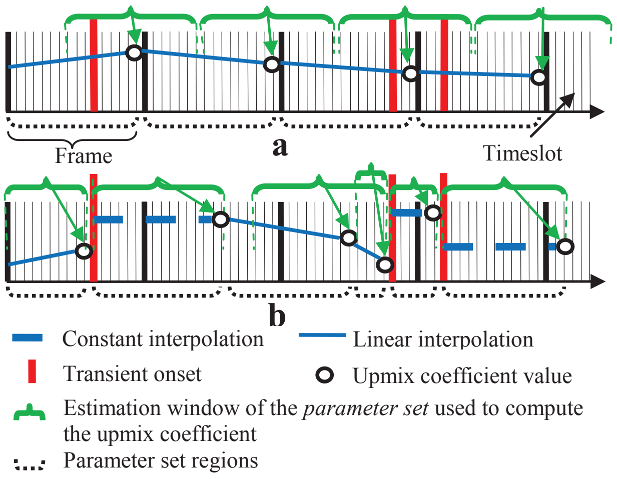

Figure 2. Parameter sets region splitting, sizing of the analysis windows for the spatial parameters and interpolation of the upmix coefficients along the timeslot into a frequency band

At the decoder, the spatial parameters P are transformed by a function f_p into upmix matrix U, with the same time-frequency resolution. Then, U is time and frequency interpolated into a matrix $\widetilde{U}$ by a function f_I in order to match the time-frequency resolution of the decoded downmix $\widehat{D}$. Finally, the upmix function f_u computes two output channels $\hat{X}_1$ and $\hat{X}_2$ from the downmix $\widehat{D}$ and the interpolated upmix matrix $\widetilde{U}$.

3. ADAPTIVE CODING IN SPATIAL PARAMETRIC CODERS

The proposed technique is to apply a strategy of adapting the size of the analysis windows for the spatial parameters, for the assignation of the parameter set and for the interpolation of the upmix coefficients based on the presence of relevant transient onsets into the downmix. Relevant transients are of high interest for spatial audio coding. Due to the precedence effect [10], transients have a strong influence on the perceived localization of following audio events, so a bad estimation of the spatial context of a transient leads to a bad localization of the transient himself and the following items (for some 100 ms depending on the signal [10]). Transients may also be associated with a sudden change in the audio scene, for example, a strong instrument entering in a spatial position highly lateralized from other instruments that are playing at this moment. In this case, changes must quickly be taken into account to avoid alteration of the transient. The alteration may be a spreading, a false localization of the transient or a combination of the two. The main reason for false localization is that the beginning of the transient may be spatialized with the spatial context of the preceding audio scene. This is partly a consequence of the smoothing of the discontinuities (typically by using a linear interpolation) that introduces delays, behaving like a low pass filter. Another reason is that the spatial parameters are not well estimated around a transient, because the window starting before the transient may spread on it and the next window may start after the onset of the transient.

The first task is to detect the transients in the downmix. Based on the onset positions, variable-size parameter sets are created, with corresponding estimation windows for the spatial parameters. At the decoder, interpolation of the upmix matrix is done differently for parameters sets regions starting on transients, and for the others. All these points are detailed in the next sub-sections and illustrated on Figure 2-b.

3.1. Transient Onset detection

The onset detection localizes the timeslots where onsets of relevant transients occur. These timeslots are then used to scale analysis windows and parameters set.

There are several methods in the literature that can be used for onset detection in musical signal [9]. Given the class of spatial parametric coders we are focused on (see Figure 1), it is more convenient to detect the onset in a time-frequency domain. For our experiments, we used the hybrid complex QMF 71 bands from MPEG [4], both for transient detection and parameter estimation and synthesis. With 48 kHz sampling frequency, the time resolution in the T/F domain is about 1.3 ms. Standard onset detection techniques working in a time-frequency domain, such as *high frequency content* and *spectral difference* [9], gave sufficient performance for our application: we do not want to detect every onsets, but only the most significant, which are the ones with a high transient-to-background power ratio.

When an onset is detected, it may be required to force an inhibition timeslot delay in order to avoid successive detections of the same transient. In our tests, we used an inhibition timeslot delay of 16 timeslots which is equivalent to 23 ms for our 48 kHz audio samples.

3.2. Parameter sets region splitting

A parameter set region corresponds to a temporal non overlapping window in the T/F domain, at witch the parameter set is assigned. Regions have a constant width N_{max} in absence of transients. Around a transient, if a region starts before and stops after its onset, it is shortened to stop just before the transient. Then, the next region starts at the transient position with a width N_{max} or just before the next transient if they are separated from less than N_{max} timeslots.

This way, the separation of the spatial context just before and during the transient can be handled correctly. The next step is to estimate the spatial parameters for each region.

3.3. Parameter sets estimation

Parameter sets are sets of spatial parameters calculated for a region. Estimation windows do not necessarily coincide with the parameter sets regions: generally, in a region, the parameter set values are affected to the last time slot of the region. This way, as soon as a parameter set and the information on its region are received, it is possible to estimate the spatialization parameters for the whole region by interpolation between the last parameter set and the received one. As the parameter set is the image of the spatial context around the last time slot, it is legitimate to center the estimation window on this timeslot. In this situation, the estimation window has the same width as the parameter set region, but shifted forward by a half width. This is what is done with our technique when no transient is present on a half width forward. In the opposite case, the estimation window is shifted forward until just before the transient, in order to avoid a bias in the estimation due to the transient. If the region starts on a transient, then the estimation window coincides with the parameter set region; it starts on the onset of the transient and has a width of N_{max} or shorter if a second onset follows the first one with a distance inferior to N_{max}. This estimation without shift forward is justified by the sudden change in the spatial context at the transient position, and by the precedence effect [10].

3.4. Temporal interpolation of the upmix matrix $\boldsymbol{U}$

Once spatial parameter regions and estimation windows are correctly positioned around transients, the corresponding adaptive interpolation must be performed to improve decoding quality of the transients.

As illustrated on Figure 1, at the decoder, down-sampled spatial parameters are used to compute an upmix matrix at the same sampling rate. It is then up-sampled by temporal and frequency interpolation to the T/F resolution of the audio signal. The number of timeslots to interpolate by parameter set region is equal to the region width minus one.

Temporal discontinuities into the upmix coefficients may lead to several audible artifacts in the decoded audio signal, often perceived as “clicks”. Linear interpolation is used in MPS and Parametric Stereo to avoid them. However, the impossibility to introduce discontinuities prevents instantaneous changes of the spatial context when a transient appears. This way, the transient may be spatially spread, leading to a bad localization and a lack of sharpness caused by the spatial spreading of its energy.

To overcome these issues, we propose an adaptive interpolation based on the presence or absence of a transient at the start of the parameter set region. When the region does not start with a transient, we apply a linear interpolation between the upmix coefficients of the last timeslot of the previous region, and the upmix coefficients of the last timeslot of the current region, as used in MPS. In the opposite case, interpolation by a constant value is used on the whole region with the upmix coefficients computed from the parameter set. In this last case, the constant interpolation leads to a discontinuity at the transient position, with a hop size corresponding to the distance between the upmix coefficient computed from the previous parameter set and the one from the current parameter set. So, the changes in the spatial context are immediately applied at the onset of the transient, leading to a precise localization and sharpness of the transient.

Again, introducing discontinuities into the upmix coefficients may lead to artifacts, but there is a masking effect of the transients when the discontinuity is introduced at the onset and when the power of the transient is sufficiently high compared to the power of the background (other audio events and stationary signals). Therefore, artifacts on the background may not be audible, as observed by the continuity illusion [11].

4. PERFORMANCE EVALUATION AND DISCUSSION

4.1. Experimental setup

As MPEG USAC was under development during the writing of this article, we chose to implement a coder and a decoder following the part 6.10 in [5]. The configuration we chose use Channel Level Difference (CLD), Inter-Channel Coherence (ICC) and Inter-channel Phase Difference (IPD) as set of spatial parameters, estimated on 28 bands, with a refresh rate of approximately 46 ms for 48 kHz samples. The core coder has been bypassed in order to evaluate only the behaviors of the spatial model. The filter bank we used for spatial analysis and synthesis is the same as the one used in USAC and MPEG Surround. Parameters have been quantized using the fine quantization tables for all parameters [4].

For our experiments, both “standard” and “proposed” versions have been done as follows. The baseline “standard” version is when constant estimation windows are used and linear interpolation is used between each upmix coefficient as on Figure 2-a. The “proposed” version is the coding technique developed here, described in section 3 and illustrated on Figure 2-b. This last one is not entirely compliant with part 6.10 in [5]. Indeed, regions of parameter sets may spread between two frames and interpolation of the upmix matrix is not systematically linear and depends on the presence of transient’s onset at the start of a parameter set region.

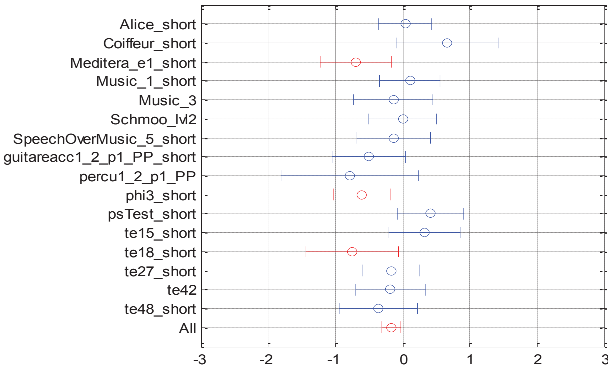

Figure 3. Mean score across the 14 listeners. Error bars indicates 95% confidence interval

4.1. Listening Test

This test employed comparisons of two impaired conditions with one reference and follows the ITU-R BS.1284-1 recommendations [8]. It includes a training phase with 3 items and a main phase with 16 items. The following versions of each item were included in the test: the original as shown reference, a USAC "baseline" version and Our coding strategy applied to the USAC "baseline".

14 expert listeners participated in the test. All excerpts were presented over headphones. The results per item, averaged across subjects and the global result averaged across items are given in Figure 3. Negative score means that the proposed technique is closer to the reference than the standard technique, and reciprocally for positive scores. This figure shows a significant improvement of the proposed technique for 3 items, but also for the mean of all items. The proposed technique provides no significant degradation for any of the items.

4.2. Algorithmic delay

Parameter set estimation requires the knowledge of whether there is a transient into the next frame. It induces an algorithmic delay of $2N_{max}$ at the encoder.

4.3. Spatial parameters bitrate

Allowing the detection of transients spaced by 16 timeslots at minimum, our technique increased by 5% the mean bitrate for the spatial parameters across the 16 audio samples of the test, with a maximum of 12% for the highly percussive sample *percu_2_p1_PP*. Notice that this 5% increase was obtained on critical samples with several relevant transients, and it is equivalent to an absolute increase of only 0.2 kbps.

Possible extensions may reduce the bitrate without affecting the improvement provided by our technique. The most relevant would be to exploit the precedence effect that can apply for delays of some 100 ms for musical signals [10]. This way, it may be that we do not have to send the parameter set following the one starting on a transient. Then, a linear interpolation can be applied between the parameter sets around the removed one. Another solution would be to increase the minimum distance between two successive onset detections, which would lead to detect fewer transients.

5. CONCLUSIONS

In this article, we presented a global strategy of spatial parameters analysis window sizing and upmix interpolation. The most relevant point we demonstrate is that introducing temporal discontinuities into the upmix coefficients at the start of the transients can improve audio quality; sharpness and localization of the relevant transients are improved. Possible artifacts resulting from discontinuities are not perceptible when the transient-to-background noise power ratio is high enough.

6. REFERENCES

[1] Jürgen Herre, « From Joint Stereo to Spatial Audio Coding - Recent Progress and Standardization », *DAFx'04 7th International Conference of Digital Audio Effects* (Naples, Italy, 2004).

[2] F. Baumgarte and Christof Faller, « Binaural Cue Coding - Part I: Psychoacoustic Fundamentals and Design Principles », *IEEE Transactions on Speech and Audio Processing* 11, n^{o}. 6 (2003): 509–519.

[3] Jeroen Breebaart and al., « Parametric coding of stereo audio », *EURASIP J. Appl. Signal Process.* 2005 (2005): 1305-1322.

[4] ISO/IEC 23003-1:2007, «MPEG-D (MPEG audio technologies), Part 1: MPEG Surround », 2007

[5] ISO/IEC 23003-3, «MPEG-D (MPEG audio technologies), Part 3: Unified Speech and Audio Coding », 2012

[6] ISO/IEC 14496-3:2005, «MPEG-4 Audio Fourth Edition, Part 3, subpart 8 (Parametric Stereo)», 2005

[7] Jeroen Breebaart and al., « Parametric coding of stereo audio », *EURASIP J. Appl. Signal Process.* 2005 (2005): 1305-1322.

[8] « ITU-R BS.1284-1 General methods for the subjective assessment of sound quality »

[9] J.P. Bello, L. Daudet and al., « A Tutorial on Onset Detection in Music Signals », *Speech and Audio Processing, IEEE Transactions on* 13, n^{o}. 5 (September 2005): 1035 - 1047.

[10] Blauert, J., « Spatial Hearing - The Psychophysics of Human Sound Localization », *MIT Press* (1983): chapter 3.1.

[11] Michael C Kelly et Anthony I Tew, « The Continuity Illusion Revisited: Coding of Multiple Concurrent Sound Sources » (MPCA-2002, Leuven, Belgium, 2002)