NOVEL LOW COMPLEXITY COHERENCE ESTIMATION AND SYNTHESIS ALGORITHMS FOR PARAMETRIC STEREO CODING

Yue Lang$^1$, David Virette$^1$, Christof Faller$^2$

Huawei European Research Center, Germany$^1$. Illusonic GmbH, Switzerland$^2$

ABSTRACT

In this paper, we present novel low complexity coherence estimation and corresponding synthesis algorithms for parametric stereo coding. Inter-channel correlation (IC) is an important parameter for parametric stereo coding as it represents the degree of similarity of the channels and is strongly related to the perception of width and diffuseness of the stereo image. It is relevant for most audio and music contents to re-generate ambience, stereo reverberation, source width, and other perceptions related to spatial impression. In the state-of-the-art parametric stereo codec, the IC estimation and corresponding synthesis algorithms are very complex which prevents their use for complexity-constrained applications. Hence, we introduce novel low complexity coherence estimation and corresponding synthesis algorithms for stereo coding. A subjective listening test shows that with the proposed algorithms, the perceived quality for very low bit rate parametric stereo is improved with a limited computational complexity cost.

Index Terms— Inter-channel correlation, parametric stereo coding, de-correlation

1. INTRODUCTION

Recent progress in low bit rate stereo audio coding has been made with several applications to audio codec targeting broadcast and streaming. Parametric stereo has been introduced in several MPEG standards (HE-AACv2 and USAC) in order to reduce the bit rate for stereo coding. Parametric Stereo (PS) coding or Binaural Cue Coding (BCC) [1-4] consist in representing the stereo signal as a mono down-mix, which is encoded with a legacy mono encoder, together with limited side information representing the stereo image with perceptual parameters. PS and BCC use those parameters (the spatial cues) to synthesize the stereo signal from the mono down-mix. The spatial cues are usually defined as:

- Inter-channel Level Difference (ILD) measuring the level difference (or balance) between channels,
- Inter-channel Time Difference (ITD) or Inter-channel Phase Difference (IPD) describing respectively the time or phase difference between channels,
- Inter-channel Coherence (IC) which represents the coherence (or amount of correlation) between channels.

A basic parametric stereo coder may use ILD as a cue needed for generating the stereo signal from the mono down-mix audio signal. Such a scheme has been introduced for conversational application in [5], showing good quality for speech content. However, this limited representation of the spatial image usually gives poor results for noisy speech or music contents. In addition, when coding binaural stereo signals e.g. for 3D audio or headphone based surround rendering, an IPD may also play a role to reproduce phase/delay differences between the channels. More sophisticated coders may also use the IC, which represents a degree of similarity between the audio channels.

IC are usually estimated in frequency domain and defined as the normalized cross-correlation coefficient after phase alignment according to the IPD. An estimation of the IC, as defined in [2], is given by

$$IC(b) = \frac{\sum_{k=k_b}^{k_e} X_L(k)X_R^*(k)}{\sqrt{\sum_{k=k_b}^{k_e} X_L^*(k)X_L(k)\sum_{k=k_b}^{k_e} X_R(k)X_R^*(k)}}$$

where $X_L(k)$ and $X_R(k)$ are the Short Term Fourier Transform (STFT) coefficients of the two channels, $*$ denotes complex conjugate, and $k_b$ is the start frequency bin of band $b$. According to this equation, IC takes its values between 0 and 1.

At the decoder side, IC synthesis may be implemented using de-correlators in frequency domain as described in [2]. However, the known estimation (at the encoder side) and synthesis (at the decoder side) approaches for multi-channel audio signals may suffer from an increased complexity (due to the cross-correlation computation and de-correlation filters) and usually require a high bit rate (IC is estimated, coded, and transmitted for every sub-band). For conversational application, the very low bit rate parametric multichannel audio coding schemes have not only the constraint on bit rate, but also some limitation on available computation power as this kind of stereo/multichannel audio codec usually targets the implementation in handsets for which long battery life is crucial. Therefore, a very low
complexity and low bit rate IC coding scheme is necessary. This paper presents an approach for coherence estimation and synthesis algorithms suitable for the application scenario mentioned above. The paper outline is as follows. First the whole structure of the parametric stereo coder which includes coherence estimation and synthesis is introduced in Section 2. Section 3 discusses a low complexity coherence estimation algorithm. Section 4 describes a new way of synthesizing the coherence between the channels, based on using simple time-domain de-correlators. Section 5 gives the subjective listening test results and analysis before concluding.

2. THE CODEC STRUCTURE

![Fig. 1. Parametric stereo audio encoder and decoder.](image)

The basic block diagram of the proposed parametric spatial audio encoder and decoder is shown in Figure 1. The boxes with dashed lines are the contribution of this paper. The stereo input signal is first processed by a parameter extraction and a down-mix module. The mono down-mix is encoded using an arbitrary mono audio coder. Usually a legacy audio encoder is used in order to offer the backward compatibility with existing mono decoder. In this paper, the mono codec is based on the ITU-T G.722 Annex B which is a super wideband extension of G.722. The extracted spatial parameters (ILD, Coherence) are quantized before being multiplexed into the bit stream together with the down-mix. The sub-band ILDs are extracted by

\[
ILD(b) = 10 \log_{10} \frac{\sum_{k} X_i(k) X_i^*(k)}{\sum_{k} X_o(k) X_o^*(k)}
\]

where \(X_i(k)\) and \(X_o(k)\) are the STFT coefficients of the two channels, * denotes complex conjugate, and \(k\) is the first frequency bin of band \(b\). In order to keep the bit rate as low as possible, a single full band inter-channel coherence parameter is quantized and transmitted.

At the decoder side, the de-multiplexer splits mono and spatial parameter information. The mono audio signal is decoded by the legacy ITU-T G.722 Annex B decoder and fed into the spatial synthesis stage, which reinstates the spatial cues based on the decoded spatial parameters. The spatial parameter extraction and synthesis is performed in a complex STFT domain on a 5 ms frame basis.

3. LOW COMPLEXITY COHERENCE ESTIMATION

The IC is usually estimated based on a normalized cross-correlation coefficient obtained after phase alignment of the input channels according to the IPD. Hence, the computation of the coherence requires first the computation of the ITD or IPD per sub-band. Moreover, the computation of a normalized cross-correlation requires more operations than a simple cross-spectrum, especially for fixed point implementation. IC relates to the degree of similarity between the channels. In a more perceptual point of view, the IC is an indication of the perceived width or diffuseness of the sound field. A decreasing IC (IC close to 0) is perceived as increasing source width until the phantom source splits into two distinct sources, one on the left side and one on the right side. An increasing IC (close to 1) represents a point audio source in the stereo image. The proposed approach does not rely on the direct computation of IC between channels, but estimates the coherence based on the decoded spatial parameters. The IC parameter represents the phase difference between two channels. When the signals in both channels are very different, the average (IPD\(_{\text{mean}}\)) of extracted IPDs in each sub-band within a frame changes quickly between two consecutive frames. Indeed, a stable stereo image with directional sources and without diffuse sources produces stable IPDs over consecutive frames. A new parameter IPD\(_{\text{dist}}\) is therefore introduced to represent this instability measure. IPD\(_{\text{dist}}\) is defined as the absolute distance between IPD\(_{\text{mean}}\) in current frame and the long term average of the \(N\) past frames local IPD\(_{\text{mean}}\) (noted IPD\(_{\text{meanLT}}\)). It can be seen that if the IPD\(_{\text{mean}}\) parameter is stable over the previous frames, the distance becomes close to 0. The distance is then equal to zero when the phase difference is stable over the time. This distance between the long term average of IPD and the local average of IPD gives a good estimation of the short-term similarity of the channels.

We found that, during a correlated segment of audio signal (for instance for speech signal), the IPD\(_{\text{dist}}\) becomes very small due to the stable phase difference between channels. During diffuse parts of the audio input (for instance for reverberated music signal), IPD\(_{\text{dist}}\) becomes much bigger and will be close to 1, if the input channels are strongly decorrelated. Based on this observation, it is then concluded that the IC (which is calculated by the state of the art methods) and IPD\(_{\text{dist}}\) have an indirect inverse relation. This relation is illustrated in Figure 2 for various stereo signals.
The proposed coherence estimation algorithm uses the similarity measure $IPD_{dist}$ to roughly estimate the coherence. The cross-spectrum requires a lower complexity than the normalized correlation calculation. Moreover, if the IPDs are already extracted in the parametric spatial audio encoder, which is the case if the encoder generates a scalable bitstream with IPD values for higher bit rates, this cross-spectrum and IPD values are already computed and the additional complexity required to compute the coherence is very limited as it is only based on simple average computations.

Estimation of the coherence includes the following steps:

- A time frequency transform (STFT) is applied to the left and right input channels,
- A cross-spectrum for each frequency sub-band is computed by
  \[ c(b) = \sum_{k=0}^{M-1} X_L(k)X_R^*(k) \]  
  where $c(b)$ is the cross-spectrum of sub-band $b$. $X_L(k)$ and $X_R(k)$ are the STFT coefficients of the two channels, $k_0$ is the start frequency bin of band $b$. For a more stable estimation of the IPD, cross-spectrum can be smoothed over frames.
- The IPDs are calculated per sub-band based on the cross- spectrum as
  \[ IPD(b) = \angle c(b) \]  
  where the operation $\angle$ is the argument operator to compute the angle of $c(b)$.
- The averaged IPD ($IPD_{mean}^{(i)}$) over the interesting frequency sub-bands is also computed by
  \[ IPD_{mean}^{(i)} = \frac{1}{M} \sum_{k=1}^{M} IPD(k) \]  
  where $i$ represents the frame index and $M$ is the number of the frequency sub-bands which are taken into account for the computation of the average.
- A long term average of the $IPD_{mean}^{(i)}$ is computed as the average over the last $N$ frames as
  \[ IPD_{meanLT}^{(i)} = \frac{1}{N} \sum_{m=0}^{N-1} IPD^{(i-m)}_{mean} \]  

In order to evaluate the stability of the IPD parameters, the distance $IPD_{dist}$ between $IPD_{mean}^{(i)}$ and $IPD_{meanLT}^{(i)}$ is computed,
\[ IPD_{dist} = |IPD_{mean}^{(i)} - IPD_{meanLT}^{(i)}| \]  
relating to the evolution of the IPD during the last $N$ frames.

In order to limit the phase wrapping effect, $IPD_{dist}$ is smoothed over two consecutive frames as
\[ IPD_{dist, sm}^{(i)} = W_{dist} \cdot IPD_{dist}^{(i-1)} + (1-W_{dist}) \cdot IPD_{dist}^{(i)} \]  
where $W_{dist}$ is the smoothing factor set to 0.9922.

Finally, the coherence parameter $C_{global}$ (for global coherence) is quantized on 2 bits according to Table 1.

<table>
<thead>
<tr>
<th>IC index</th>
<th>$IPD_{dist, sm}^{(i)}$ region</th>
<th>$C_{global}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0.67 \leq IPD_{dist, sm}^{(i)}$</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>$0.52 \leq IPD_{dist, sm}^{(i)} &lt; 0.67$</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>$0.36 \leq IPD_{dist, sm}^{(i)} &lt; 0.52$</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>$IPD_{dist, sm}^{(i)} &lt; 0.36$</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: global coherence $C_{global}$ mapping table

4. LOW COMPLEXITY TIME DOMAIN DECORRELATION

Figure 3 summarizes the processing of the proposed parametric stereo synthesis scheme. The G.722 Annex B decoded mono downmix signal $x(n)$ is converted to a short-time spectral representation by STFT, denoted $X(k)$. The processing for one parametric stereo parameter band is shown in detail in the figure. All other bands are processed similarly. Scale factors $W_L$, $W_R$, and $W_D$ are applied to the frequency representation of the downmix signal $X(k)$ to generate the frequency representations of the left correlated
sound $Y_L(k)$, right correlated sound $Y_R(k)$, and left-right uncorrelated sound $D(k)$ respectively.

The generated frequency representation of the three signals $Y_L(k)$, $Y_R(k)$ and $D(k)$ are converted back to the time domain by using an inverse STFT. Two independent decorrelators ($D_1$ and $D_2$ on the figure) are applied to $d(n)$ to generate two (ideally) independent signals, which are added to $y_L(n)$ and $y_R(n)$ to generate the final stereo output left and right signals $z_L(n)$ and $z_R(n)$.

The calculation of $w_L$, $w_R$, and $w_D$ depends on the amplitude of down-mix signal. If the amplitude of the down-mix signal is defined as

$$|M| = g_{D}\sqrt{|L|^2 + |R|^2}$$

At the decoder, based on the ILD, the relative powers of the left and right channels are calculated in the following way,

$$P_L(b) = \frac{1}{10+10^{ILD(b)/20}}, \quad P_R(b) = \frac{10}{10+10^{ILD(b)/20}}$$

where $b$ is the index of sub band.

Given the global coherence, the amount of diffuse sound in the left and right channels, $P(b)$ is computed similarly as shown in [7],

$$P_L(b) + P_R(b) - \sqrt{(P_L(b) + P_R(b))^2 - 4(1-C_{global})P_L(b)P_R(b)}$$

$$P(b) = \frac{1}{2}(P_L(b) + P_R(b)) - \sqrt{(P_L(b) + P_R(b))^2 - 4(1-C_{global})P_L(b)P_R(b)}$$

(Note that in the following, for brevity of notation, we often omit the indices $b$ and $k$.

Before using further, $P_D$ is lower bounded by zero and upper bounded by the minimum of $P_L$ and $P_R$. The scale factors are computed such that the resulting three signals $Y_L$, $Y_R$, and $D$ have power equal to $P_L$, $P_R$, and $P_D$, i.e.

$$w_L = \frac{P_L - P_D}{\sqrt{2}g_D P}$$

$$w_R = \frac{P_R - P_D}{\sqrt{2}g_D P}$$

$$w_D = \sqrt{\frac{2}{g_D P}}$$

where the power of the downmix is $P = 1$ (since $P_L$, $P_R$, and $P_D$ are normalized, see above) and the factor of $g_D$ relates to the normalization that is used for the downmix input signal. In the conventional case, when the downmix is the sum multiplied by 0.5, $g_D$ is then chosen to be 0.5.

If the amplitude of the down-mix signal is

$$|M| = \frac{|L| + |R|}{2}$$

Some adaptations need to be made. The ILDS are applied to the downmix at the decoder side using the following formula for $c_1$ and $c_2$

$$c_1 = \frac{2c}{1+c} = \frac{2|L|}{|L| + |R|}$$

$$c_2 = \frac{2}{1+c} = \frac{2|R|}{|L| + |R|}$$

where

$$c = 10^{\frac{ILD}{20}} = \frac{|L|}{|R|}$$

Those definitions of $c_1$ and $c_2$ allow recovering the correct amplitude for the left and the right channel. $P_L$, $P_R$ and $P_D$ are still defined according to the previous definition (10) and (11).

If we define a case where $C_{global} = 1$, and the amplitude of the down-mix signal is defined as (13). We also used the definition of $P_L$, $P_R$ and $P_D$ and apply them on the downmix signal, we would then have

$$|\hat{R}| = w_R |M| = \frac{P_R}{g_D} |M|$$

$$|\hat{R}| = 2\sqrt{\frac{|L|^2 + |R|^2}{|L|^2 + |R|^2}}$$

To cancel the effect of the mismatch between downmix computation and assumption on $P_L$ and $P_R$, some adaptations are needed.

If we define

$$d = 10^{\frac{ILD}{20}} = \frac{|L|^2}{|R|^2}$$

we have

$$g = \frac{1 + d}{1 + c} = \frac{|L|^2}{|L|^2 + |R|^2}$$

For the downmix defined as (13), the $w_L$, $w_R$, and $w_D$ are adapted to keep the energy of the left and right channel according to:

$$w_L = 2\sqrt{(P_L - P_D) \cdot g}$$

$$w_R = 2\sqrt{(P_R - P_D) \cdot g}$$

$$w_D = 2\sqrt{P_D \cdot g}$$

In the case $C_{global} = 1$, those $w_L$, $w_R$, and $w_D$ definitions allow us to obtain exactly the same result as with the weighting factor $c_1$ and $c_2$.

A very low complexity way of doing de-correlation is to simply use different delays for $D_1$ and $D_2$. Moreover, those delays allow simple implementation in the frequency domain which is shown in Figure 4, as they are a multiple of the 5 ms frame size, the decorrelation is simply obtained by a delay line on STFT coefficients (as 2 and 4 frames delay are necessary for $D_1$ and $D_2$ respectively).

Fig 4. Frequency domain implementation of coherence synthesis

Moreover, those delays allow simple implementation in the frequency domain which is shown in Figure 4, as they are a multiple of the 5 ms frame size, the decorrelation is simply obtained by a delay line on STFT coefficients (as 2 and 4 frames delay are necessary for $D_1$ and $D_2$ respectively).
Implementation in frequency domain keeps the complexity low as it does not require an additional inverse STFT compared to normal ILD synthesis.

5. SUBJECTIVE LISTENING TEST RESULTS

In order to evaluate the subjective quality of the proposed parametric stereo coding scheme, a Ref/A/B test using the 7 grading scales which is defined in [9] was conducted to compare the quality of the proposed algorithm with a stereo codec similar to [5]. Both codecs have a bit budget of 40 bits per frame for the stereo parameters (including ILD and IC when available). As explained in section 3, a single coherence parameter ($C_{global}$) per frame was quantized with 2 bits and transmitted when necessary. One bit was used to indicate to the decoder whether the coherence parameter was present or not. $IPD_{dist\_sum} < 0.2$ indicates that the coherence is close to 1 and there is no need to transmit the coherence parameter. Compared to the codec similar to [5], the bits were “stolen” to the ILD quantization (1 or 3 bits per frame).

In total, 16 test items were used. They are selected from four categories (ITU-T music categories): classical orchestral, classical vocal, modern instrumental and modern vocal. Four items are used in each category. They were all pre-filtered to the 50-14000 Hz bandwidth and normalized in level. 8 expert listeners participated in the subjective listening test, using high-quality headphones.

The test results are shown in Figure 5. In the horizontal axis, 1 to 16 are the test items and 17 represents the average result over all items. The mean score and corresponding 95% confident interval are shown on the figure. We can see from the figure that, for the music, the quality of proposed coherence estimation and synthesis technique is better than the reference codec.

We also checked the quality of clean and noisy speech. 16 binaural clean speech items were used for testing. The outputs of the proposed codec and the reference codec were identical, which means that for most of the clean speech content, it is expected that the proposed algorithm will not affect the quality. Two types of noise were considered in the noisy speech test: office noise (20dB) and interfering talker noise (15dB). For speech with office noise, the average SSSNR (segmental signal to noise ratio) of the 8 used items is 89dB which indicates that the coherence synthesis is very rarely activated. For speech with interfering talker, in 7 out of 8 used items were identical, and the SSSNR for the last one was 93.36dB.

From subjective and objective test results, we can see that the proposed algorithm improves the quality of stereo music items without affecting the quality of stereo speech.

6. CONCLUSION

In this paper, we proposed a novel parametric stereo coding scheme for low bit rate and low complexity applications. This coding scheme includes new coherence estimation and synthesis algorithms which offer a good complexity/quality tradeoff for constrained applications. The combination of the coherence estimation and synthesis in a low complexity stereo codec for conversational application has shown good performances for music signals without any degradation for speech content.

7. REFERENCES