Master Thesis: Parametric Coding for Spatial Audio

Author: Bertrand Fatus
Supervisor at Orange: Stéphane Ragot
Supervisor at KTH: Sten Ternström

July-December 2015
Abstract

This thesis presents a stereo coding technique used as an extension for the Enhanced Voice Services (EVS) codec [10] [8]. EVS is an audio codec recently standardized by the 3rd Generation Partnership Project (3GPP) for compressing mono signals at chosen rates from 7.2 to 128 kbit/s (for fixed bit rate) and around 5.9 kbit/s (for variable bit rate). The main goal of the thesis is to present the architecture of a parametric stereo codec and how the stereo extension of EVS may be built. Parametric stereo coding relies on the transmission of a downmixed signal, sum of left and right channels, and the necessary audible cues to synthesize back the stereo image from it at the decoding end. The codec has been implemented in MATLAB with use of the existing EVS codec.

An important part of the thesis is dedicated to the description of the implementation of a robust downmixing technique. The remaining parts present the parametric coding architecture that has been adapted and used to develop the EVS stereo extension at 24.4 and 32 kbit/s and other open researches that have been conducted for more specific situations such as spatial coding for stereo or binaural applications.

Whereas the downmixing algorithm quality has been confronted to subjective testing and proven to be more efficient than any other existing techniques, the stereo extension has been tested less extensively. Yet the quality reached with the proposed reconstruction algorithms tends to highlight the potential of the codec that could be revealed by future work.
Frequently used abbreviations and notations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR</td>
<td>Adaptative Multi-Rate.</td>
</tr>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project.</td>
</tr>
<tr>
<td>BCC</td>
<td>Binaural Cue Coding.</td>
</tr>
<tr>
<td>EVS</td>
<td>Enhanced Voice Services.</td>
</tr>
<tr>
<td>FB</td>
<td>Full Band (20 - 20000 Hz).</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform.</td>
</tr>
<tr>
<td>HRIR</td>
<td>Head related impulse response, modeling time transformation of sound from a source to left and right ear entrances.</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head related transfer function, modeling frequency transformation of sound from a source to left and right ear entrances.</td>
</tr>
<tr>
<td>ICC</td>
<td>Inter-channel coherence, i.e. degree of similarity between channel signals.</td>
</tr>
<tr>
<td>ICLD</td>
<td>Inter-channel level difference, i.e. level difference between channel signals.</td>
</tr>
<tr>
<td>ICTD</td>
<td>Inter-channel time difference, i.e. time difference between channel signals.</td>
</tr>
<tr>
<td>ILD</td>
<td>Inter-aural level difference, i.e. reception level difference between the ears. ILD might sometimes be used for ICLD to lighten the notations.</td>
</tr>
<tr>
<td>IPD</td>
<td>Inter-channel phase difference, i.e. phase difference between channel signals.</td>
</tr>
<tr>
<td>ITD</td>
<td>Inter-aural time difference, i.e. reception time difference between the ears.</td>
</tr>
<tr>
<td>MDCT</td>
<td>Modified discrete cosine transform.</td>
</tr>
<tr>
<td>MDST</td>
<td>Modified discrete sine transform.</td>
</tr>
<tr>
<td>MUSHRA</td>
<td>MUltiple Stimuli with Hidden Reference and Anchors.</td>
</tr>
<tr>
<td>NB</td>
<td>Narrow Band (300 - 3400 Hz).</td>
</tr>
<tr>
<td>OPD</td>
<td>Overall phase difference.</td>
</tr>
<tr>
<td>QMF</td>
<td>Quadrature mirror filter.</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform.</td>
</tr>
<tr>
<td>SWB</td>
<td>Super Wide Band (50 - 14000 Hz).</td>
</tr>
<tr>
<td>USAC</td>
<td>Unified Speech and Audio Coding.</td>
</tr>
<tr>
<td>WB</td>
<td>Wide Band (100 - 7000 Hz).</td>
</tr>
</tbody>
</table>
## Contents

1 Introduction 9
  1.1 Background and motivations 9
  1.2 Thesis outline 9
  1.3 Key concepts of parametric stereo/multichannel coding 10

2 Downmix 13
  2.1 Basic downmix 13
    2.1.1 Passive downmix : AMR-WB+ downmix 13
    2.1.2 Passive downmix with scaling : e-AAC+ downmix 14
    2.1.3 Phased downmix : Samsudin and al. downmix 14
  2.2 USAC downmix 15
  2.3 G.722 D downmix 17
  2.4 Adami and al’s downmix 18
  2.5 More on the sum signal phase 19
    2.5.1 Degenerated phase of the sum signal 19
    2.5.2 Finding the issue with phase computation of the downmix 20
    2.5.3 ISD : an indicator to delimit the problematic situation 22
    2.5.4 Linear extrapolation from region 1 23
    2.5.5 Extrapolation from region 2 24
    2.5.6 Phase continuity between successive frames 25
  2.6 A new hybrid downmix 26
    2.6.1 Introducing the hybrid downmix 26
    2.6.2 Subjective test at 10 ms frame 29
  2.7 Optimal downmix 32
    2.7.1 Introducing the hybrid+ downmix 32
    2.7.2 Subjective tests at 20 ms frame 36
  2.8 Robustness of the hybrid+ downmix 38

3 Parametric stereo coding 41
  3.1 Cue averaging 41
  3.2 Cue quantization 43
  3.3 Basic upmix 45
  3.4 Ambience synthesis 46
    3.4.1 Ambience through coherence 46
    3.4.2 Ambience through reverberation 49

4 EVS stereo extension at 32 kbit/s 51
  4.1 EVS dual mono 51
  4.2 EVS stereo extension 52
    4.2.1 Asymmetrical EVS stereo extension 52
    4.2.2 Symmetrical EVS stereo extension 53

5 Spatial coding for stereo application 55
  5.1 Source localization 55
  5.2 ITD estimation 57
  5.3 ILD estimation 59
6 Spatial coding for binaural application
   6.1 Virtual 3D sound ....................................................... 61
   6.2 From Stereo to 3D ....................................................... 62

7 Conclusion ................................................................. 65

8 Appendix A : Sound sample treatment before test .................. 67

References ................................................................. 69
1 Introduction

1.1 Background and motivations

For numerous applications such as cellphone telecommunication systems, low cost transmission and storage of audio signals are essential. In order to meet the capacity and coverage requirements of those systems lossless transmission techniques are not suitable thus leading to the development of lossy coding techniques. These techniques, focusing on perceptual models, aim at removing perceptual irrelevancies so that the audio coding can achieve greater compression efficiency without degrading the audio quality.

Audio codecs, resulting of the combination of an encoder and a decoder, usually have to face a very restrictive compromise between bit rate allocation for transmission, algorithmic complexity, delay requirements and audio quality. Those requirements are often defined for 4 inherent qualities of audio signals associated with the bandwidth available: narrow band (NB, 300-3400 Hz), wide band (WB, 100-7000 Hz), super wide band (SWB, 50-14000 Hz) and full band (FB, 20-20000 Hz). For stereo transmission the bit rates range from 16 kbit/s to 128 kbit/s in order to meet the capacity of the systems and deliver the best quality along with a maximum delay restriction corresponding to the human sensibility around 40-50 ms.

For stereo signals, those codecs follow three main approaches: dual mono coding, mid side coding or parametric stereo coding. Dual mono coding is the simplest architecture where a stereo signal is seen as the sum of two independent entities: the left and the right channels. Those two channels are encoded and decoded regardless of any correspondence between the two. Mid side coding aims at encoding and decoding the sum and the difference of the two input channels. Usually left and right channels have a lot of common information, so encoding the sum is as expensive as encoding one channel, but the difference can contain less information and so the bit rates used can be asymmetrically distributed over the two to enhance the resulting quality after reconstruction. A more complex version of mid side coding with prediction can also be found in method such as [24]. Finally parametric stereo coding intends to separate the signal content and its spatial distribution. As for mid side coding, the whole content of a stereo signal can be found in the sum of the two channels, but instead of transmitting the difference, this method extract perceptual cues to be transmitted to synthesize back the spatial image of the original file. Methods such as stereo intensity coding use a similar approach.

1.2 Thesis outline

The goal of this thesis is to develop a parametric stereo codec as an extension of the existing monaural EVS codec. EVS has been developed for speech and music coding from narrow band to full band at multiple native bit rates and also adapted to the adaptive multi-rate (AMR) codecs already existing modes to enhance compatibility. A short description of useful concepts to understand stereo coding is made at the end of this introduction. An important first part (section 2) of the report is then dedicated to the research and development of a robust summing method of the two input channels. The second part (sections 3 and 4) deals with the decoding process and the synthesis of the stereo output. The last part of the report (sections 5 and 6) highlights open researches that have been conducted during the internship and might prove useful for future work.
1.3 Key concepts of parametric stereo/multichannel coding

The idea behind parametric multichannel coding is to maximize the compression of a multichannel signal by transmitting parameters describing the spatial image. For stereo input signals the compression process basically follows one idea: synthesizing one signal from the two input channels and extract parameters to be encoded and transmitted in order to add spatial cues for the synthesized stereo at the receiver’s end. This can be illustrated as a block diagram:

![Figure 1: The idea behind parametric coding](image)

The parameter estimation is made in the frequency domain with consecutive short time frequency transforms (STFT, MDCT) or in sub-band domain (complex QMF, etc.). The spectra of the signal segments \( X_i[k] \) are then divided into sub-bands \([0, 1, ...k_b, ...N_b]\) used to average relevant parameters and compress even more the encoded signal. For stereo applications, the most useful parameters are the following:

The inter-channel level difference (ICLD) defined as:

\[
\text{ICLD}[b] = 10 \log_{10} \left( \frac{\sum_{k=k_b}^{k_b+1-1} X_1[k]X_1^*[k]}{\sum_{k=k_b}^{k_b+1-1} X_2[k]X_2^*[k]} \right)
\]  

(1.1)

In that case the ICLD can be seen as a parametric inter-aural level difference (ILD) at the channel level. In a practical sense the ILD represents the relative loudness difference perceived by our ears when the emitting sound source is not situated in the front position (if a speaker is standing on the left side of a listener, the latter will receive more sound in his left ear and ILD > 0 dB with our definition).

The ICC defined as the inter-channel coherence, which can be computed as:

\[
\text{ICC}[b] = \frac{\left| \sum_{k=k_b}^{k_b+1-1} X_1[k]X_2^*[k] \right|}{\sqrt{\left( \sum_{k=k_b}^{k_b+1-1} X_1[k]X_1^*[k] \right) \left( \sum_{k=k_b}^{k_b+1-1} X_2[k]X_2^*[k] \right)}}
\]  

(1.2)

The ICC is linked to the perceptual volume of a sound source: the greater the ICC the smaller the sound source appears.
The inter-channel phase difference (IPD) defined as:

\[
\text{IPD}_b = \angle \left( \sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k] \right) \quad (1.3)
\]

The overall phase difference (OPD) is the phase difference between one channel and the downmixed signal and is defined as:

\[
\text{OPD}_b = \angle \left( \sum_{k=k_b}^{k_{b+1}-1} X_1[k]M^*[k] \right) \quad (1.4)
\]

![Figure 2: IPD and OPD geometrical representation](image)

In common downmix techniques the redundant parameter OPD is used. The OPD can be expressed using the other inter channel parameters:

\[
\text{OPD}_k = \angle (c[k] + \text{ICC}[k]e^{j\text{IPD}[k]})
\]

\[
c[k] = 10^{\text{ICLD}[k]/20} \quad (1.5)
\]

![Figure 3: OPD geometrical representation](image)

As long as the ICC is equal to 1, meaning that the signals are correlated, the OPD can be exactly expressed as:

\[
\text{OPD}_k = \arctan \left( \frac{c_2[k] \sin(\text{IPD}[k])}{c_1[k] + c_2[k] \cos(\text{IPD}[k])} \right) \quad (1.6)
\]

where:

\[
c_1[k] = \sqrt{\frac{c^2[k]}{1 + c^2[k]}}, \quad \text{and} \quad c_2[k] = \sqrt{\frac{1}{1 + c^2[k]}} \quad (1.7)
\]
Yet when the ICC is not equal to 1, this definition leads to the wrong OPD and can then cause degradation of audio quality.

To tackle that problem Hyun et al. developed an algorithm using a new parameter called PDS (phase difference sum). The idea of Hyun et al. is to use both the OPDs from the left and right channel as:

\[ \text{CPD1}[k] = \angle \left( \sum_{k=k_b}^{k_{b+1}-1} X_1[k]M^*[k] \right), \quad \text{and} \quad \text{CPD2}[k] = \angle \left( \sum_{k=k_b}^{k_{b+1}-1} M[k]X_2^*[k] \right) \tag{1.8} \]

Then they recompute the IPD using the total phase difference PDS, which leads to a better reconstruction of the IPD when \( ICC < 1 \) [14].

\[ \text{PDS} = \text{CPD1} + \text{CPD2} = \angle \left( \left( c + \frac{1}{c} \right) ICC + 2 \cos(\text{IPD}) ICC^2 \right) e^{j\text{IPD}} + (1 - ICC^2) \tag{1.9} \]

When used in STFT the IPD can be translated into the time domain through the inter-channel time difference (ICTD) as simple conversion using the center frequency of each band and the size of the frame used for the transformation \( N_{fft} \). The ICTD related to the time envelope is unique for each frame and its conversion to IPD is proportional to the frequency bin as:

\[ \text{IPD}[k] = 2\pi \frac{f_s k}{N_{fft}} \cdot \text{ICTD} \tag{1.10} \]

Similar to the ICLD, the ICTD has a perceptual meaning. The path lengths taken by a propagating sound to reach our ears will be different when the emitting source is not situated in the median plan. Thus a time delay, the inter-aural time difference, will occur and help us evaluate the position of that source in space.
2 Downmix

As depicted in the introduction, the stereo signal needs to be downmixed to one single signal \( m[n] \), \( M \) on the figures) containing as much information as possible in order to be used as the starting point of the synthesis. Numerous downmixing techniques exist and some will be presented in this part. They basically follow at least one of the two ideas:

- Find a way to get the best average of all the channels.
- Use one channel as a reference and ensure that it contains sufficient information to reproduce the other one.

All the downmixing methods will be presented for a stereo input with two channels: Left \( (x_1[n], L \) on the figures) and Right \( (x_2[n], R \) on the figures). Some of those methods can be extended to multichannel input signals, however that topic will not be presented in this paper.

2.1 Basic downmix

2.1.1 Passive downmix: AMR-WB+ downmix

The simplest way of creating a downmixed signal is to compute a direct average of the two input channels either in time domain or in frequency domain:

\[
m[n] = \frac{x_1[n] + x_2[n]}{2} \quad M[k] = \frac{X_1[k] + X_2[k]}{2}
\]

(2.1)

The main problem of this downmix is that the amplitude of the sum signal \( m \) is strongly affected by the phase relation between \( x_1 \) and \( x_2 \). In the worst case of two identical phase opposed signals, the sum signal will be zero. Furthermore, the amplitude variations of the sum signal related to the phase relation between the input channels will lead to strong degradation of the quality of the downmix.

This method in time domain is the one used in the 3GPP AMR-WB+ codec [2].
2.1.2 Passive downmix with scaling: e-AAC+ downmix

In order to avoid some energy loss when input channels are combined, a scaling gain can be used to ensure that the energy of the downmixed signal is the same as the sum of the inputs. The stereo scale factor can be computed in the sub-band domain as (the double indexing \([k,n]\) is linked to the complex QMF method):

\[
\gamma_{k,n} = \sqrt{\frac{|X_1[k,n]|^2 + |X_2[k,n]|^2}{0.5|X_1[k,n] + X_2[k,n]|^2}} \tag{2.2}
\]

Then the sum signal is computed as before and equalized with this gain:

\[
M_{k,n} = \frac{X_1[k,n] + X_2[k,n]}{2} \cdot \gamma_{k,n} \tag{2.3}
\]

Yet in case of out of phase inputs the gain \(\gamma[k]\) tends to overestimate the actual energy and local restrictions have to be made, such as limiting its value to 2 (i.e. 6dB) [23]. This method is used in the 3GPP e-AAC+ codec [3].

2.1.3 Phased downmix: Samsudin and al. downmix

To tackle the out of phase issue, one solution is to align the signals before summing them. Thus the IPD can be used to rotate \(X_2[k,n]\) in the direction of \(X_1[k,n]\) by computing:

\[
M_{k,n} = \frac{X_1[k,n] + X_2[k,n]}{2} \cdot e^{j \text{IPD}[k]} \tag{2.4}
\]

![Figure 5: Phase alignment](image)

The resulting sum signal is also automatically preserving the energy of the sum of the original channels [23]. The major problem in this method is the IPD determination which can lead to errors when averaged over all the frequency bins of a sub-band. As depicted above, the IPD is generally calculated per sub-band with (1.3). So if the consecutive bins are uncorrelated, meaning that \(X_1[k]X_2[k]\) and \(X_1[k+1]X_2[k+1]\) have random phases, that can lead to cancellation between the terms as displayed below:
2.2 USAC downmix

Following the same idea as Hyun et al.’s, Kim et al. developed their algorithm [15] taking into account the phase jump issue that can occur when the input signals are oscillating around an out of phase position. As depicted in the figure below, when $X_1$ and $X_2$ are out of phase, the sum signal $M$ created using the OPD and IPD can be affected by strong phase shifts from one frame to the next leading to cancelation of the sum signal on its own:

To solve this problem, the energy of one or both of the channel signals is adapted into an offset (ofst) so that the difference between two consecutive OPDs is reduced as displayed on figure (8).

By confronting OPD differences with a step size $Q_{\text{step}}$, the algorithm adapts the relative energy of both channels to avoid phase jumps of the sum signal. The approach is described in [15].

A new OPD is computed such as:

$$\text{OPD}_{\text{new}}[k] = \arctan \left( \frac{c_2[k] \sin(\text{IPD}[k])}{c_1[k] \cdot \text{ofst} + c_2[k] \cos(\text{IPD}[k])} \right)$$  \hspace{1cm} (2.5)
Figure 8: Decreasing the energy of $X_2$ to reduce OPDs variations between two consecutive frames

This new OPD is calculated when the phase difference in subsequent frames is greater than the acceptable value $Q_{step}$, ofst is then calculated to reduce the phase difference to that reference value:

$$\arctan\left(\frac{c_2[k] \sin(IPD[k])}{c_1[k] \cdot \text{ofst} + c_2[k] \cos(IPD[k])}\right) = Q_{step}$$

Leading to:

$$\text{ofst} = \frac{c_2[k]}{c_1[k]} \left(\frac{\sin(IPD[k])}{\tan(Q_{step})} - \cos(IPD[k])\right)$$

The original sum signal given by:

$$M[k] = e^{-jOPD[b]} \cdot X_1[k] + e^{-(jOPD[b] - IPD[b])} \cdot X_2[k]$$

Is changed into a more stable one:

$$M'[k] = e^{-jOPD_{new}[b]} \cdot X_1[k] + e^{-(jOPD_{new}[b] - IPD[b])} \cdot X_2[k]$$

As their authors explain in [18]: the purpose of this technique is to handle anti-phase signals by applying an unbalanced weighting of the left and right channels during downmixing and upmixing process.
2.3 G.722 D downmix

In order to reduce the bit costs, Huawei technologies developed a coding scheme using whole band parameters instead of numerous sub-band parameters [26] [1]. The whole band IPD is averaged over the lowest frequency sub-bands IPD (for instance $a = 2$ to $c = 6$):

$$\text{IPD}_{WB} = \frac{1}{E} \sum_{k=a}^{k=c} e^{(i)}[k] \text{IPD}[k]$$  \hspace{1cm} (2.10)

where $e^{(i)}[k] = X_1[k]X_1^*[k] + X_2[k]X_2^*[k]$, $E = \sum_{k=a}^{k=c} e^{(i)}[k]$ \hspace{1cm} (2.11)

In similar ways, the whole band Inter-channel time difference (ICTD) and ICC are computed using enhanced algorithms to ensure a meaningful averaging process over the targeted sub-bands.

The whole band parameters are also used to transmit parameters and avoiding phase cancelations. This particular use of the whole band IPD and ICTD is presented in [26]. The downmixed signal is then computed as:

$$\angle M[k] = \angle X_1[k] - \frac{1}{1 + c[k]} \left( \text{IPD}[k] - \text{IPD}_{WB} \right)$$

$$|M[k]| = \frac{|X_1[k]| + |X_2[k]|}{2}$$  \hspace{1cm} (2.12)

When the left channel is dominant, $c[k] \to \infty$ and $\angle M[k] = \angle X_1[k]$, on the contrary if the right channel is dominant then $c[k] \to 0$ and $\angle M[k] = \angle X_1[k] - \left( \text{IPD}[k] - \text{IPD}_{WB} \right) = \angle X_2[k] + \text{IPD}_{WB}$ and in that case a decision tree is supposed to have reset IPD$_{WB}$ to 0. When the IPD is close to $\pi$ the whole band IPD becomes $\pi$ (phase wrapping is applied so that if one subband IPD is equal to $-\pi$ it is set to $\pi$) and so the sum signal phase is set to the phase of the left channel, thus avoiding phase jumps and cancelation.

All in all the complete downmixing process along with the multiple decisions trees are described in details in [1]. As several known critical signals are used to enhance this algorithm and ensure its reliability, it is probable that for new critical signals not taken into account the algorithm will fail. Its robustness will be tested later in the report (see subsection 2.7.2 and 2.8.2).
2.4 Adami and al’s downmix

In a different perspective, this downmix method detailed in [4] aims at using one channel signal as a reference $X_1$ and suppressing its coherent components within the other one before downmixing. $X_2$ is assumed to be the sum of a coherent part $W \cdot X_1$ and uncorrelated signal part $U$ with respect to $X_1$. Thus $X_2$ can be expressed as:

$$X_2[k] = W[k] \cdot X_1[k] + U[k]$$  \hspace{1cm} (2.13)

Then the downmixed signal is created from a rescaled version of $X_1$ to keep the right energy and the extracted uncorrelated signal $U$:

$$M[k] = G_{X_1}[k] \cdot X_1[k] + U[k]$$  \hspace{1cm} (2.14)

In theory a suppression gain $\hat{G}$ is estimated to suppress the correlated part of the signal component within $X_2$ so that the uncorrelated signal estimate $\hat{U}$ can be added to the rescaled reference signal $X_1$ using an estimated gain factor $\hat{G}_{X_1}$. Thus the estimate of the desired downmixed signal is given by:

$$\hat{M}[k] = \hat{G}_{X_1}[k] \cdot X_1[k] + \hat{G}[k] \cdot X_2[k]$$  \hspace{1cm} (2.15)

As this approach uses estimators based on previous frames of STFT, the smoothing process is directly embedded into the computation of the downmixed signal. Nonetheless problems occur for signals containing several transitions. As the method works with estimations, when the signal is too transient then the estimation fails and the downmix is of poor quality.
2.5 More on the sum signal phase

2.5.1 Degenerated phase of the sum signal

The phase of the sum signal is generally at the heart of the downmix issues. Let us consider a simple downmix as defined in section 2.1. The phase of the sum signal is not well defined for IPDs close to \( \pm \pi \). As presented in the figure below, the phase of the sum signal is degenerated when the IPD is equal to \( \pi \) or \(-\pi\).

![Error in the sum signal phase for passive downmix](image1.png)

Figure 11: Error in the sum signal phase for passive downmix

In practice one would want a linear relation between the phase of the monaural signal \( \angle M \) and the IPD so that the decoding process would not lead to ambiguity which could translate into reconstruction errors. If the phase of the sum signal was well defined then no degeneration problems should occur around large IPDs and \( \angle M \) should look like :

![Theoretical \( \angle M \)](image2.png)

Figure 12: Phase of the sum signal for downmixes if no degeneration occured
2.5.2 Finding the issue with phase computation of the downmix

In order to simplify the problem, we can consider 3 situations:

Situation 1: \( |\text{IPD}[k]| < \frac{\pi}{2} \) or both signals are in phase and no problem of downmix phase definition occurs for any ILDs:

![Figure 13: Small IPDs and accurate definition of sum signal phase (left), corresponding region 1 on the original plot (right)](image)

Situation 2: \( |20 \log(\text{ILD}[k])| >> 0 \) or one signal is much stronger than the other and no problem of downmix phase definition occurs as the sum signal aligns with the dominant signal:

![Figure 14: Large ILDs and accurate definition of sum signal phase (left), corresponding region 2 on the original plot (right)](image)

In both those situations a basic downmix as defined in (2.1) is sufficient, yet in order to ensure a good energy conservation, one would prefer:

\[
\angle M[k] = \angle \left( \frac{X_1[k] + X_2[k]}{2} \right),
\]

\[
|M[k]| = \frac{|X_1[k]| + |X_2[k]|}{2}
\]  \hspace{1cm} (2.16)
Situation 3: None of the previous conditions are met and the sum signal phase is ill defined:

Figure 15: Large IPDs and comparable strengths of input signals: out of phase cancelation (left), corresponding region 3 on the original plot (right)

In this last situation, using (2.16) will lead to wrong sum signal phase estimation. One way to tackle this problem is to use non degenerated situations (situation 1 or 2) to extrapolate the exact phase value for this region. Two interpolating methods are presented in subsection 2.6.4 and 2.6.5. Subsection 2.6.3 presents an estimator that can be used to delimit situation 1 and 2 from situation 3 and thus apply changes only when necessary.
2.5.3 ISD : an indicator to delimit the problematic situation

In order to distinguish situation 1 and 2 from situation 3, an indicator can be created; called here the inter spectral difference (ISD). This indicator can be used for bin to bin evaluation (ISD\([k]\)) or subband evaluation (ISD\([b]\)):

\[
\begin{align*}
\text{ISD}[k] &= \frac{|X_1[k] - X_2[k]|}{|X_1[k] + X_2[k]|}, \\
\text{ISD}[b] &= \frac{1}{(k_{b+1} - k_b)} \sum_{k=k_b}^{k_{b+1}-1} \frac{|X_1[k] - X_2[k]|}{|X_1[k] + X_2[k]|} 
\end{align*}
\] (2.17)

A graphic representation of the ISD helps understanding how it works :

![Figure 16: Values taken by the ISD for different ILDs and IPDs](image1)

If the IPD values are within \([-\pi/2, \pi/2]\) (situation 1) then the ISD is automatically lower than 1 (blue lines). On the other hand if one of the channel is strongly dominant (the ILD is far from 0 dB) then the ISD value shrinks down to 1 (situation 2). Only for absolute values of IPD around \(\pi\) and comparable signal strengths is the ISD value likely to explode. Thus by using a threshold such as 1.3 (limiting the relative error of the sum signal phase to 12 %) we can separate situation 1 and 2 from 3 using the ISD :

![Figure 17: Separating situation 1 and 2 from 3 using the ISD indicator](image2)
### 2.5.4 Linear extrapolation from region 1

In this subsection we will present how to extrapolate the wanted phase of the sum signal in region 3 from region 1. A similar extrapolation system can be found in [13]. The idea is to return to a situation where the IPD is small and then go back to the actual large IPD value. In a simplified configuration ($\angle L = 0$) this translates into a linear regression over one ILD line.

Let us consider one situation, for instance ILD = 3.8 dB and IPD = 3 (point A). As we can see from the figure below we are in region 3 and the phase of M is ill defined. So we are going to retract back to region 1 where $IPD_{\text{ref}} = \pi/4$ (point B) and then extrapolate the exact phase for the signal M (point C).

![Figure 18: Extrapolation from region 1](image)

In order to do the extrapolation we can consider the equation of the straight line drawn in red on the graph, whose angle is equal to $\alpha$:

$$\angle M = \alpha \cdot IPD$$ (2.18)

The $\alpha$ coefficient can be computed using the value at the origin $(0,0)$ in our case and the value at the reference taken : $IPD_{\text{ref}} = \pi/4$. But as we are back in region 1, we can use the actual value given by the basic downmix for $\angle M_{IPD_{\text{ref}}}$

$$\alpha = \frac{\angle M_{IPD_{\text{ref}}}}{IPD_{\text{ref}}} = \frac{\angle (|X_1| + |X_2|e^{-jIPD_{\text{ref}}})}{IPD_{\text{ref}}},$$ (2.19)

As we can see, this formulation also takes into account the ILD and so the $\alpha$ coefficient is implicitly computed for the right ILD. This coefficient being known we can then unfold back to our extrapolated version at IPD:

$$\angle M = \frac{\angle (|X_1| + |X_2|e^{-jIPD_{\text{ref}}})}{IPD_{\text{ref}}} \cdot IPD.$$ (2.20)
2.5.5 Extrapolation from region 2

The idea behind G.722 D algorithm is to avoid those phase error in the sum signal. So in order to recreate the right phase, they actually interpolate its value using the barycenter of the two input signals as:

\[
\angle M[k] = \frac{\angle X_1[k] \cdot |X_1[k]| + \angle X_2[k] \cdot |X_2[k]|}{|X_1[k]| + |X_2[k]|}
\]

Which rewritten leads to:

\[
\angle M[k] = \frac{1}{1 + |X_2[k]|/|X_1[k]|} \cdot \text{IPD}[k]
\]

The phase representation is then correct for all the values of the IPD. This can be represented as below: starting from point A we look at the maxima of ILD for such an IPD, when \(|X_1| >> |X_2|\) (point B) and \(|X_2| >> |X_1|\) (point B’) and evaluate \(\angle M\) there (i.e. \(\angle X_1\) or \(\angle X_2\)). Then we apply ILD weights on those values and find the barycenter according to the dark dotted line on C which corresponds to the exact phase of the sum signal.

Figure 19: Extrapolation from region 2
2.5.6 Phase continuity between successive frames

Another issue concerning the downmix phase is that even though an algorithm such as depicted above are used to avoid phase degeneration inside an analysis frame, the continuity of the phase for consecutive frames is not guaranteed. If the left and right channels are out of phase, jumps of the IPD from $-\pi$ to $\pi$ between frames will result in jumps of the monaural signal phase from $-\pi/2$ to $\pi/2$ and lead to cancelation between successive frames (see figure 7). USAC idea as presented in subsection 2.2 is to rescale one of the channel in order to avoid such phase jumps. In Huawei algorithm, the whole band IPD which is typically equal to $-\pi$ in case of out-of-phase channels allows to avoid the phase jump by giving $\angle M = \angle L$.

Another method to avoid phase jumps is to translate the problem to $\pm \pi/2$ instead of $\pm \pi$. In fact a phase jump of $\pi$ from one frame to the next for the monaural signal is problematic because it will lead to a complete cancelation of the bin in the overlap-add region. If this jump is reduced to $\pi/2$ then the auto cancelation is only partial and the downmix quality is better. One way to implement this new phase definition of the monaural signal is to use the mid and side signals defined as:

$$\text{Mid}[k] = \frac{X_1[k] + X_2[k]}{2}$$
$$\text{Side}[k] = \frac{X_1[k] - X_2[k]}{2} \quad (2.23)$$

Then using the threshold ISD = 1 delimiting $|\text{IPD}[b]| < \pi/2$ and $|\text{IPD}[b]| > \pi/2$ regions, we define two corresponding values for $\angle M$:

$$\angle M[k] = \angle \text{Mid}[k] \quad \text{when ISD}[b] \leq 1$$
$$\angle M[k] = \angle \text{Side}[k] \quad \text{when ISD}[b] > 1 \quad (2.24)$$

The new phase of the monaural signal can be found as represented below:

![Figure 20: New phase assignment to the monaural signal using Mid and Side vectors (for different ILD values)](image)
We can see that for all the situations there is no more phase jumps of the monaural signal around $\pi$ and $-\pi$. The only problematic case is if the $X_2$ signal is evolving around $-\pi/2$ or $\pi/2$ and is dominant, as the monaural phase would align with it, jumps of amplitude $\pi$ can occur when switching from Mid to Side.

Testing this algorithm on a basis of chosen signals, the results are mostly good for critical signals where important phase oppositions occur, yet for simpler signals the phase jump at $\pm\pi/2$ creates strong audible artefacts. In light of that, this idea was not pushed forward.

2.6 A new hybrid downmix

2.6.1 Introducing the hybrid downmix

In light of all that has been discussed in subsection 2.5, a new hybrid downmix has been implemented in order to fulfill all of the requirements presented: limiting the phase jumps around $\pm\pi$, keeping the energy ratio, avoid out of phase cancelation and ensure a good quality for non critical signals. A threshold of 1.3 was kept as it showed the best potential for all the signal basis after listening evaluation. This threshold is used to separate the different situations and can be correlated to an IPD value of $7\pi/12$. The maximum of the ISD for a given IPD is always given for $|X_1| = |X_2|$, and for IPD = $7\pi/12$ that maximum is equal to 1.3.

![Figure 21: ISD and IPD for varying ILDs](image)

Looking at the linear monaural signal phase and the reconstruction from a passive approach, we can evaluate the maximum relative error using a threshold of 1.3 and implementing a passive downmix up to that value. This is the same as computing the downmix with a passive method for all ILDs and a maximum IPD of $7\pi/12$. If we use a passive downmix for ISD < 1.3, the relative error on the monaural signal phase for IPD below $7\pi/12$ and all possible values of ILD doesn’t exceed 19% (see figure 22). Yet it is possible that for IPD larger than $7\pi/8$ ILDs above a certain range can create a larger error. The worst case is for IPD = $\pi$ (see figure 16), IPD = $\pi$ correspond to the highest red curve). There is the ISD dependence on the ILD the strongest, and even if there is a dominant signal, if the ILD drops, the threshold is reached. For that worst scenario, the actual maximum relative error created by using a passive downmix is of 12 % (see figure 23).
Figure 22: Maximum error of 19% on the monaural phase signal using an ILD independent passive downmix with threshold $\text{ISD} = 1.3$ (i.e. $\text{IPD} = 7\pi/12$).

Figure 23: Maximum error of 12% on the monaural phase signal using a passive downmix below $\text{ISD} = 1.3$ and an exact construction above for $\text{IPD}$ kept at $\pi$ and varying ILD.

A first new downmix will be implemented using a perfect phase implementation given by eq (2.25) above $\text{ISD} = 1.3$ and a gain factor such as (2.2) will be added to ensure energy conservation. We will refer to this downmix as switched G722/passive:
if ISD[b] > 1.3 : for \( k \in [k_b : k_{b+1} - 1] \)

\[
\angle M[k] = \frac{\angle X_1[k] \cdot |X_1[k]| + \angle X_2[k] \cdot |X_2[k]|}{|X_1[k]| + |X_2[k]|} \\
|M[k]| = \frac{X_1[k] + X_2[k]}{2} \cdot |\gamma[k]| 
\]  

(2.25)

(2.26)

if ISD[b] \leq 1.3 : for \( k \in [k_b : k_{b+1} - 1] \)

\[
M[k] = \frac{X_1[k] + X_2[k]}{2} \cdot |\gamma[b]| 
\]  

(2.27)

Yet for numerous listening tests, the results is not as good as expected. For stereo inputs where left and right channels are slightly correlated, implementing the theoretical average phase between left and right loses any meaning and the result is a degradation of the quality of the downmix. Thus changing the phase of a signal even for creating a perfect bijective downmix might not be the best approach. So using a simpler algorithm, another downmix called \textit{switched Samsudin/passive} has been implemented as :

\[
\text{if ISD}[k] > 1.3 : \\
\angle M[k] = \angle X_1[k] \\
|M[k]| = \frac{X_1[k] + X_2[k]}{2} \cdot |\gamma[b]| 
\]  

(2.28)

(2.29)

\[
\text{if ISD}[k] \leq 1.3 : \\
M[k] = \frac{X_1[k] + X_2[k]}{2} \cdot |\gamma[b]| 
\]  

(2.30)

This algorithm is inspired by Samsudin’s, yet using the ISD we eliminate the failing case where the phase is aligned with the one of \( X_1 \) even though \( X_1 \) is noise and so not relevant. When the ISD is larger than 1.3 both signals are comparable in strength and so the downmix signal is phased with an actual sound signal.

As the phase of the downmix can be the most sensitive parameter for audio quality, two approaches have been used as hybrid downmixes and only the best of them (chosen by listening comparison) has been kept. For the first hybrid downmix the ISD is computed subband by subband ISD[b] and all the frequency bins inside that subband are subjected to the same decision : either passive downmix or a downmix similar to G722. For the second hybrid downmix the decision is made bin by bin : ISD[k].
2.6.2 Subjective test at 10 ms frame

In order to evaluate the quality of several downmixes presented in the previous subsections and the new ones developed, a subjective test has been conducted. This test followed a modified MUSHRA methodology, signals are randomly ordered and blindly presented and listeners have to evaluate their quality on a scale of 0 (poor quality) to 100 (excellent quality). To do so a stereo reference is openly available and an anchor is hidden in the subjective test for each file, which differs from an actual MUSHRA test. The anchor was created using a passive downmix degraded by random modulation of the amplitude and the phase. All the signals are encoded into 16-bit PCM with sampling rate 16 kHz.

For each sample under test there is one open reference and as many signals to be tested as algorithms (in our case 7) plus the anchor. The reference is presented as “reference” and the 8 signals are presented in random order using letters from A to H, all can be listen to whenever the subject want to. As the downmix quality is not easily comparable with the reference in stereo, subjects were asked to try and evaluate the relative quality with respect to the content of the stereo signal: how well the relative loudness of right and left channels contents were kept in the downmix, how well the downmix was keeping that content at a good quality, how well the overall signal was enjoyable and free of audible artefacts (clics, jumping phases, etc.). Portions of the signals could be selected and played in loop for specific comparison between signals under test. During this process the listener can grade and change at will that grading for all the different downmixed signals. When the listener is happy with his final score distribution, he can save it and move on to the next sample under test.

The downmixes kept for the tests are the following:

- **G722D**: G722 D standard, used with the C code
- **Adami&al**: Matlab implementation taken from the Fraunhofer and F.A. Universität patent [5]
- **USAC**: Matlab implementation taken from their review [15]
- **Samsudin**: Matlab implementation taken from their review [23]
- **Hybrid1**: Matlab implementation of our downmix as presented in the previous subsection (*switch G722/passive*)
- **Hybrid2**: Matlab implementation of our downmix as presented in the previous subsection (*switch Samsudin/passive*)
- **MDCT**: Matlab implementation the Hybride2 downmix in MDCT domain
In order to be representative and also to evaluate the robustness of the downmixes, several sample tests have been used. The samples are presented in Table 1 with their characteristics in order to highlight the stressed features of the downmixes.

<table>
<thead>
<tr>
<th>Nr</th>
<th>Item</th>
<th>Signal Class and features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unfo</td>
<td>Singing and music with IPD evolving around $\pi$</td>
</tr>
<tr>
<td>2</td>
<td>Songout</td>
<td>Music with synthesized OPD effects</td>
</tr>
<tr>
<td>3</td>
<td>Noise</td>
<td>One channel is noise</td>
</tr>
<tr>
<td>4</td>
<td>Music21</td>
<td>LR music with one instrument per channel</td>
</tr>
<tr>
<td>5</td>
<td>Guitare</td>
<td>Music with repeated short attacks</td>
</tr>
<tr>
<td>6</td>
<td>Salvation</td>
<td>Broadcast chorus music</td>
</tr>
<tr>
<td>7</td>
<td>Twinkle</td>
<td>Voice over music</td>
</tr>
<tr>
<td>8</td>
<td>Canyon</td>
<td>Outdoors movie like sounds</td>
</tr>
<tr>
<td>9</td>
<td>Jazz</td>
<td>Broadcast jazz music</td>
</tr>
<tr>
<td>10</td>
<td>Music_3</td>
<td>LR music</td>
</tr>
<tr>
<td>11</td>
<td>Speech_AB</td>
<td>AB microphone simulation with clean speech</td>
</tr>
<tr>
<td>12</td>
<td>Noisy_AB</td>
<td>AB microphone simulation with noisy speech</td>
</tr>
<tr>
<td>13</td>
<td>Speech_MS</td>
<td>MS microphone simulation with clean speech</td>
</tr>
<tr>
<td>14</td>
<td>Noisy_MS</td>
<td>MS microphone simulation with noisy speech</td>
</tr>
<tr>
<td>15</td>
<td>Movvoice_BI</td>
<td>Binaural voice anechoic</td>
</tr>
<tr>
<td>16</td>
<td>Bombarde_BI</td>
<td>Binaural pipe music with noise</td>
</tr>
</tbody>
</table>

Table 1: Sound file basis under test.

Half a dozen participants took this test and the results were roughly analyzed in order to simply highlight the strengths and weaknesses of our new downmixing methods. Some of the participants were well trained and used to subjective evaluation testing while other were untrained. Both their results were interesting because the actual quality of the downmixes was better evaluated by the trained listeners, but many artefacts picked up by them weren’t audible or at least perceived as a nuisance for the rest of the testers. As the ultimate listener will be an untrained one in a large majority, it is helpful to use this feedback to try and tackle first the strong degradations pointed out by all the testers and then work more specifically on small artefacts often unnoticed by the untrained majority. The results are presented on figure 24.
As we can see, our hybrid2 downmix has the best average quality over all the samples. Yet this is partly due to the fact that it never actually crashes on one particular signal. The Adami&al downmix has a really good quality for all the signals except two: music21 and guitare. Also we can highlight that our hybrid2 downmix doesn’t seem to be optimized for signals with large IPD differences such as AB or binaural type signals.
2.7 Optimal downmix

2.7.1 Introducing the hybrid+ downmix

As the results proved in the first round of testing (see figure 24), the new algorithm hybrid2 was promising yet it was lacking some robustness for signals such as AB types, binaural or music and unfo. After looking more closely to those signals, we highlighted that they where all featuring large phase differences. Due to the wrapping effect of the STFT on the phase, especially for higher bins, the ISD might not be discriminating enough because no reference of phase is taken for each frame. We developed a new enhanced hybrid2 downmix, hybrid+, that could deal with those problematic types of signals and smooth inter frame effects.

From the first test we can see that the Samsudin downmix has good results for the problematic signals, and as our hybrid2 already uses it, we decided to switch to a complete Samsudin downmix for those signals. In order to do so we found that the realigned ICC (ICCr) computed over the whole frame (frame index : \( f \)) was a good indicator to switch to this downmix:

\[
\text{ICCr}[f] = \frac{\left| \sum_{k=1}^{k_{\text{max}}} X_1[k]X_2[k]e^{j\text{IPD}[k]} \right|}{\sqrt{\left( \sum_{k=k_b}^{k_b+1} X_1[k]X_1^*[k] \right) \left( \sum_{k=k_b}^{k_b+1} X_2[k]X_2^*[k] \right)}}; \quad k_{\text{max}} = \frac{N_{\text{FFT}}}{2} + 1 \quad (2.31)
\]

ICCr represents the degree of correlation between left and right channels without taking into account phases shifts. Thus for clean AB and binaural signals, the ICCr takes values close to 1, whereas for signals such as music21 and unfo_noise where left and right channels are independent, the ICCr value drops. For tested noisy AB and binaural signals, the drop in ICCr is not significant enough to crash the downmix (SNR \( \approx 10\text{dB} \)). As this drop is significant and given that signals with no correlation between left and right channels can usually be downmixed with a passive downmix, we also chose to switch, appropriately, to a passive downmix for those values. Thus we had two thresholds, a high threshold, e.g. at 0.6, and above it a Samsudin downmix, a low threshold, e.g. at 0.4, and below it a passive downmix and for values in the middle, the hybrid2 downmix which actually is an intra frame mix of Samsudin downmix and passive downmix.

![Figure 25: Strong correlation for speech AB: Samsudin downmix](image)
In order to smoothen the interframe variations of the ICCr[f] values and avoid strong transitions we added a smoothing process defined as:

$$\text{ICCr}[f] = 0.5\text{ICCr}[f] + 0.25\text{ICCr}[f - 1] + 0.25\text{ICCr}[f - 2]$$  \hspace{1cm} (2.32)

Other smoothing methods could be applied, this one was efficient enough and kept.
Finally a cross fade is applied between all the regions in order to smoothen even more the transitions. If we refer to the monaural signals from the three downmixes as: **Samsudin**: $M_s[f]$, **hybrid**: $M_h[f]$ and **passive**: $M_p[f]$. A weighted coefficient $\rho[f]$ was used defined as:

$$\rho[f] = \cos^2 \left( \frac{\pi}{2} \frac{|\text{ICCr}[f]-0.5|}{0.1} \right)$$ (2.33)

if $\text{ICCr}[f] \leq 0.4$ :

$$M[f] = M_p[f]$$ (2.34)

if $0.4 < \text{ICCr}[f] \leq 0.5$ :

$$M[f] = (1 - \rho[f])M_p[f] + \rho[f]M_h[f]$$ (2.35)

if $0.5 < \text{ICCr}[f] \leq 0.6$ :

$$M[f] = (1 - \rho[f])M_s[f] + \rho[f]M_h[f]$$ (2.36)

if $\text{ICCr}[f] > 0.6$ :

$$M[f] = M_s[f]$$ (2.37)

![Figure 28: Weights applied to the downmixes with respect to the ICCr value](image-url)
One last issue from this method is that when the left and right channels are correlated the downmixed signal is aligned with the left channel, even though the right channel might be more suitable. In the case of \textit{music\textunderscore 3}, there is a large sound area recorded and the music starts from the left and ends on the right. Yet with our method, as left and right channels are correlated, the downmixed signal all be aligned with the left and some artefacts can be heard. A dominance indicator (SGN) was added to the algorithm to be able to follow the actual dominant channel in case of correlated sound. Such a change of reference can not be used in a region where the \textbf{Samsudin} or the \textbf{hybrid2} downmix are used because that would lead to artefacts due to jumps from left phase reference to right phase reference. So only when evolving in the \textbf{passive} downmix area can this change occur without leading to degradation. An indicator taking values +1 for left reference or -1 for right reference was added and changed only when the signal would see dominance changes in the passive region. The dominance indicator was defined as:

\begin{equation}
\text{SGN}[f] = \text{sign} \left( \sum_{k=1}^{k_{\text{max}}} |X_1[k]| - \sum_{k=1}^{k_{\text{max}}} |X_2[k]| \right)
\end{equation}

Figure 29: Phase alignment reference change for \textit{music\textunderscore 3} (in black). The ICC\textsubscript{r} is multiplied by the indicator (“signed”) on this figure to better grasp the architecture of the downmix.

The SGN indicator is applied to define the reference in the \textbf{Samsudin} and \textbf{hybrid2} part of the downmix only if ICC < 0.3

\begin{equation}
\angle M[k] = \left( \frac{1 + \text{SGN}[f]}{2} \right) \angle X_1[k] + \left( \frac{1 - \text{SGN}[f]}{2} \right) \angle X_2[k]
\end{equation}

Otherwise the default choice is SGN\textsubscript{f} = 1. Thus the change occurs when the \textbf{Samsudin} is not considered for the downmix (ICC < 0.4), and will be effective when the ICC reaches again the region where that downmix is actually used, hence avoiding artefacts.
2.7.2 Subjective tests at 20 ms frame

For this second session of subjective tests, the frame length used was adapted to the new EVS standard being 20 ms. However as the G722D downmix was embedded in the G722D codec, it was kept at 10 ms. In order to simplify the test and make it less tiresome, only the algorithms showing the best results from the first test were kept, that is: G722D, Adami&al and Samsudin. Adding our hybrid+ downmix and the anchor they were only 5 downmixes to evaluate on the 16 sound files.

From all the candidates (L) who did the test we computed the mean over all the samples (N=16 × L) and the 95% confidence interval (CI) related to those values. The standard deviation \( \sigma \) over all the listenings of the same downmix is defined as:

\[
\sigma^2 = E[X^2] - (E[X])^2 = E[X^2] - \mu^2
\]  

The 95% CI represents the probability at 95% that a new tester will evaluate the downmix in that interval. The estimation of that interval is based on the hypothesis that the results from the test follow a Gaussian (or normal) distribution. That interval is calculated by:

\[
I_{95} = \left[ \mu - 1.96 \frac{\sigma}{\sqrt{N}} ; \mu + 1.96 \frac{\sigma}{\sqrt{N}} \right]
\]

The results for the 4 downmixes texted are the following:

<table>
<thead>
<tr>
<th>Downmix</th>
<th>95% CI</th>
<th>Mean</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor</td>
<td>1.19</td>
<td>8.92</td>
<td>7.73</td>
<td>10.11</td>
</tr>
<tr>
<td>Adami&amp;al</td>
<td>3.18</td>
<td>71.20</td>
<td>68.02</td>
<td>74.38</td>
</tr>
<tr>
<td>G722D</td>
<td>3.79</td>
<td>63.77</td>
<td>59.98</td>
<td>67.55</td>
</tr>
<tr>
<td>Hybrid+</td>
<td>1.62</td>
<td>80.83</td>
<td>79.21</td>
<td>82.45</td>
</tr>
<tr>
<td>Samsudin</td>
<td>3.26</td>
<td>72.80</td>
<td>69.54</td>
<td>76.06</td>
</tr>
</tbody>
</table>

Figure 30: Evaluation of the downmixes, mean in red and 95% CI in black
The new **hybrid+** downmix has been evaluated as the most robust one as over all the samples its mean mushra score is the highest and furthermore its 95% CI is not even reached by the other downmixes. The strongest feature of the **hybrid+** downmix is indeed its robustness. When comparing mushra scores samples by samples with the other algorithms, it is sometimes challenged over the actual quality of the downmix (see *Movvoice_BI*) yet it is the only downmix that never crashes on a particular sample:

This downmix, **hybrid+**, is the one that will be used for the experimental EVS stereo codec that will be presented in the remaining sections.
2.8 Robustness of the hybrid+ downmix

In order to evaluate the robustness of the hybrid+ downmix, several tests and studies have been conducted with different test signals. The tests were mainly based on subjective evaluation of the downmix for non critical signals. The studies aimed at choosing and creating critical signals considering the features of our downmix algorithm.

Using different sound files taken from the MPEG, USAC, and other associates, we run the algorithm and checked that the downmixes were of good quality. With the sound basis we had, we encountered no problematic situations.

We studied the downmix on several speech situations using different recording techniques (MS, AB, binaural, with (up to SNR=10 dB) and without noise) and the conclusion is that the realigned correlation remains high for all cases. Also the reference choice for the phase of the downmix doesn’t impact the quality of the downmix. The default reference is the left channel, yet for signals starting with people talking on the right, either the ICCr is low enough at the beginning to switch the reference, or it keeps the left channel as a reference, in any case the quality of the downmix remains good. Furthermore, in case of people talking at the same time but different positions, as once again the ICCr value stays high, there is no switch in reference throughout the conversation and the quality of the downmix is not altered.

The figure above displays the ICCr for the BinauralSpeechSeparation file which features a binaural recording of a speaker on the left, then an other on the right then both talking at the same time. As the ICCr doesn’t reach values below 0.3, the kept reference is the left channel. For frame numbers between 0 and 500 this choice is the best one for the talker is actually on the left. For frame numbers between 500 and 1000 however the talker is on the right, yet as we are in a region where the whole downmix is phased it is dangerous to operate a reference switch, nevertheless the quality of the downmix re-
mains good. Finally for frame numbers above 1000, both talkers are speaking at the same
time, yet the reference is still kept on the left channel and the quality of the downmix is
maintained. This last region could become a problematic one if the SGN indicator would
change for higher ICCr values than 0.3. Indeed we would encounter several back and forth
switching from left to right and the quality would be degraded.

If we apply an ICCr value threshold of 0.8 to allow switch references, then the downmix
follows better the actual dominant channel and for speech signal this can be useful to
increase the quality:

![Figure 32: ICCr threshold value at 0.8, multiple reference switch (in black: 1 for left, -1
for right)](image)

As we can see the first 500 frames are aligned mostly with the left channel, then when
the talker is on the right (500 to 1000 frames) the reference is changed to the right channel
and at the end when both speakers are talking several switches occur. This doesn’t affect
the quality of the downmix and even enhances it. Such use of a moving threshold for the
ICCr value could be useful.

Nonetheless using and ICCr threshold value of 0.8 for the SGN change on other several
files proved that the quality is not audibly enhanced and that it creates clicks on the
downmix signal which on the whole degrades the quality. For a general use of the hybrid+
downmix, it is preferable to apply the reference switch in the passive downmix area to
avoid artefacts, yet for particular use of the algorithm, one could choose a higher ICCr
threshold value such as 0.7 or 0.8 to better follow the phase of the actual dominant signal
without leading to important degradation of the downmix signal.
3 Parametric stereo coding

Numerous approaches for parametric stereo coding exist and rely on the transmission of the cues such as ICLD, IPD and ICC which are used in the upmix to "respatialize" the sound. In order to lower the bit rate for these cues, averaging and quantizations are applied in way to ensure the best compromise between bit costs and quality.

3.1 Cue averaging

The IPD, ICLD and ICC cues described in the introduction are given by subband. The ICLD and IPD cues could be computed for each frequency bin, yet the transmission costs would explode. So subbands are used to average those values and lower the bit cost as given in the introduction. An important aspect is that those subbands can be chosen so that the impact on the audio quality would be lessen to a minimum. As depicted in [7], the use of the Equivalent Rectangular Bandwidth (ERB) scale translates that idea. The ERB are an approximation of the human auditory filters, and as such they correspond to the critical band regions. A more complete description can be found in [16].

The ERB scale is given by:

\[ BW[f] = 24.7(0.00437f + 1) \] (3.1)

Choosing a center frequency \( f_c \), we can then average the cues over all the frequencies falling inside the corresponding bandwidth \( BW[f_c] \) (\([f_c - BW[f_c]/2, f_c + BW[f_c]/2\] translated if need be so that all bands are adjacents). But the choice of \( f_c \) are limited and defined by this of the Fourier transform. So we only need to compute each bandwidth for each frequency bin and every time a frequency bin falls into the previously frequency bin’s ERB, a subband is created gathering all those frequency bins until the next one out of the masking region is reached. Those subbands delimited by the calculated frequency bins \( k_b \) are used to average our cues:

\[
\begin{align*}
\text{ICLD}[b] &= 10 \log_{10} \left( \frac{\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_1^*[k]}{\sum_{k=k_b}^{k_{b+1}-1} X_2[k]X_2^*[k]} \right) \\
\text{ICC}[b] &= \frac{|\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k]|}{\sqrt{\left( \sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_1^*[k] \right) \left( \sum_{k=k_b}^{k_{b+1}-1} X_2[k]X_2^*[k] \right)}} \\
\text{IPD}[b] &= \angle \left( \sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k] \right)
\end{align*}
\]

In our case, as we are using a 40 ms frame length at a 16 kHz sampling frequency (i.e. \( N_{fft} = 640 \)) to be able to synthesize 20 ms frame after overlap add, we can reduce the number of transmitted cues from 321 bins to only 35 ERB subbands. Those subbands are given in the next table.
<table>
<thead>
<tr>
<th>Band number</th>
<th>ERB Frequency range (Hz)</th>
<th>Bin number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-35</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>36-57</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>58-90</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>91-126</td>
<td>4-5</td>
</tr>
<tr>
<td>5</td>
<td>127-167</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>168-210</td>
<td>7-8</td>
</tr>
<tr>
<td>7</td>
<td>211-259</td>
<td>9-10</td>
</tr>
<tr>
<td>8</td>
<td>260-314</td>
<td>11-12</td>
</tr>
<tr>
<td>9</td>
<td>315-373</td>
<td>13-14</td>
</tr>
<tr>
<td>10</td>
<td>374-439</td>
<td>15-17</td>
</tr>
<tr>
<td>11</td>
<td>440-513</td>
<td>18-20</td>
</tr>
<tr>
<td>12</td>
<td>514-594</td>
<td>21-23</td>
</tr>
<tr>
<td>13</td>
<td>595-683</td>
<td>24-27</td>
</tr>
<tr>
<td>14</td>
<td>684-784</td>
<td>28-31</td>
</tr>
<tr>
<td>15</td>
<td>785-895</td>
<td>32-35</td>
</tr>
<tr>
<td>16</td>
<td>896-1017</td>
<td>36-40</td>
</tr>
<tr>
<td>17</td>
<td>1018-1152</td>
<td>41-46</td>
</tr>
<tr>
<td>18</td>
<td>1153-1303</td>
<td>47-52</td>
</tr>
<tr>
<td>19</td>
<td>1305-1471</td>
<td>53-58</td>
</tr>
<tr>
<td>20</td>
<td>1472-1655</td>
<td>59-66</td>
</tr>
<tr>
<td>21</td>
<td>1656-1860</td>
<td>67-74</td>
</tr>
<tr>
<td>22</td>
<td>1861-2088</td>
<td>75-83</td>
</tr>
<tr>
<td>23</td>
<td>2089-2339</td>
<td>84-93</td>
</tr>
<tr>
<td>24</td>
<td>2340-2617</td>
<td>94-104</td>
</tr>
<tr>
<td>25</td>
<td>2618-2925</td>
<td>105-117</td>
</tr>
<tr>
<td>26</td>
<td>2926-3268</td>
<td>118-130</td>
</tr>
<tr>
<td>27</td>
<td>3269-3647</td>
<td>131-145</td>
</tr>
<tr>
<td>28</td>
<td>3648-4065</td>
<td>146-162</td>
</tr>
<tr>
<td>29</td>
<td>4066-4530</td>
<td>163-181</td>
</tr>
<tr>
<td>30</td>
<td>4531-5046</td>
<td>182-201</td>
</tr>
<tr>
<td>31</td>
<td>5047-5615</td>
<td>202-224</td>
</tr>
<tr>
<td>32</td>
<td>5616-6247</td>
<td>225-249</td>
</tr>
<tr>
<td>33</td>
<td>6248-6947</td>
<td>250-277</td>
</tr>
<tr>
<td>34</td>
<td>6948-7721</td>
<td>278-307</td>
</tr>
<tr>
<td>35</td>
<td>7722-8000</td>
<td>308-321</td>
</tr>
</tbody>
</table>

Figure 33: Subband table
3.2 Cue quantization

Following that idea of bit rate reduction, quantizing the IPD and ICLD cues along with taking into account audible rendering through ERB averaging, expectations can be met.

First of all, it is acknowledged that IPD over 2 kHz is not perceptible for complex signals, and that limit can even be lowered down to 1.6 kHz for pure tones [6]. So the transmission of IPD can be limited to the first 20 subbands (up to 1.6 kHz) and actually 15 in case of most signals (up to 1 kHz).

Secondly as our hearing capacities are limited, cues can be quantized without audible degradation. This is proven to be an efficient way of reducing bit rate for transmitting those cues. Codebooks with different bit allocations are used to quantize IPD and ICLD. For the IPD we use a uniform scale quantizer with 3, 4 or 5-bit codebook given by:

for 5-bit:
\[ \text{Codebook}(i) = i \cdot \frac{\pi}{16}; \quad i \in [0, 31] \] (3.2)

for 4-bit:
\[ \text{Codebook}(i) = i \cdot \frac{\pi}{8}; \quad i \in [0, 15] \] (3.3)

for 3-bit:
\[ \text{Codebook}(i) = i \cdot \frac{\pi}{4}; \quad i \in [0, 7] \] (3.4)

For the ICLD however we use an absolute non uniform 5-bit codebook for the first subband, then a relative non uniform 4-bit codebook up to 1.6 kHz then 3-bit for the rest. Those 3 codebooks are given below:

<table>
<thead>
<tr>
<th>i</th>
<th>Codebook(i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1.9</td>
</tr>
<tr>
<td>8</td>
<td>1.6</td>
</tr>
<tr>
<td>9</td>
<td>1.3</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
<tr>
<td>11</td>
<td>0.8</td>
</tr>
<tr>
<td>12</td>
<td>0.6</td>
</tr>
<tr>
<td>13</td>
<td>0.4</td>
</tr>
<tr>
<td>14</td>
<td>0.2</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>28</td>
<td>40</td>
</tr>
<tr>
<td>29</td>
<td>45</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

Figure 34: 5-bit ICLD quantization codebook
The ICC from the 6th to the last subband can be quantized with a 3-bit codebook:

\[
\text{Codebook} = [0.05, 0.15, 0.3, 0.5, 0.7, 0.8, 0.9, 1]
\]  \quad (3.5)

In conclusion, the ICC, ICLD and IPD cues can be transmitted using various codebooks over specified subband groups. Compromises between the rate allocated to the downmix and the rate allocated for spatial cues for best audio quality will be the subject of the next sections.
### 3.3 Basic upmix

In order to synthesize stereo at the decoder, ICLD and IPD are used through their
direct definitions. As the IPD is the phase difference between the left and the right
channel, a simple rotation in the Fourier domain is needed to recreate the stereo phases.
Applying the IPD to the right or the left channel is the same because only the phase
difference will have an audible effect. Yet as our downmix is by default aligned with the
left channel, in our case keeping the exact same phase for the left channel at the upmix
is more logical.

\[
\hat{X}_1[k] = M[k] \quad (3.6)
\]
\[
\hat{X}_2[k] = M[k]e^{-j\text{IPD}[b]}
\]

In the same way, the normalized ICLD is used to recreate the stereo relative loudness:

\[
c[b] = 10^{\text{ICLD}[b]/20} \quad (3.7)
\]
\[
c_1[b] = \sqrt{\frac{2c^2}{1 + c^2}}
\]
\[
c_2[b] = \sqrt{\frac{2}{1 + c^2}}
\]

The relative loudness factors are applied to the downmix:

\[
\hat{X}_1[k] = c_1[b]M[k] \quad (3.8)
\]
\[
\hat{X}_2[k] = c_2[b]M[k]e^{-j\text{IPD}[b]}
\]

Those two effects combined recreate both the delay and the relative loudness between
the two output channels, thus leading to good stereo synthesis. Yet this method is not
sufficient for complex signals such as music or multiple speakers. This synthesis is linked
to the hypothesis that the sound source is unique and punctual. For a single talker, for
instance on a cellphone conversation, this might be enough to recreate sufficient stereo
effects. But for a rich stereo input, such a synthesis would project all the sounds in one
particular direction regardless of the spatial importance of each source. For instance if
we apply this algorithm to an orchestra recording, either all the sounds will be considered
as similar to the solo player and as such all the instruments will be heard as if gathered
around that soliste in one particular direction is space, or on the contrary the solo player
will be synthesized as faded into the unlocalized rest of the orchestra.

Taking into account differences between localized sound and ambience sound is a more
complex process detailed in the next subsection. Nonetheless as such a basic method can
be sufficient, and also quite good if the cues are precise enough (bin cues instead of
subband cues), some work for the EVS stereo extension codec will be made with this
scheme.
3.4 Ambience synthesis

Ambience synthesis is a more open field of synthesis algorithms. This aspect takes into account the wideness of the sound sources as depicted in the introduction and also the background noise present in the recordings. The ICC cue is often used to synthesize ambience through different algorithms, but other techniques such as adding virtual reverberation can be found.

3.4.1 Ambience through coherence

Often the ICC is used along with a decorrelated signal generated from the downmix. The decorrelated signal is used to enrich the downmix, and the ICC is used to mix both signals. The more the original signals are correlated (the closer ICC is to 1), the less decorrelated signal will be taken into account to the stereo synthesis and vice versa. One mixing can be found in details [14] and is described in this subsection.

The first requirement of that method is the synthesis of a decorrelated signal, or orthogonal signal. One way to do so is to convolve the impulse response of a Schroeder-phase complex filter with the downmix signal [12] ($N_s = 640$, two frames):

$$h_d[n] = \frac{2^{k=N_s/2}}{N_s} \sum_{k=0}^{N_s} \cos \left( \frac{2\pi kn}{N_s} + \frac{2\pi k(k-1)}{N_s} \right)$$  (3.9)

The convolution process with the downmix signal creates the decorrelated signal $m_d[n]$. Then the mixing is embedded in an upmix matrix which can be split into 3 more meaningful matrices P, L and U. P being the phasing process and L the leveling one such as described in 3.3. U is the actual mixing using the ICC and can be seen as a scaled rotation of the $(M,M_d)$ basis prior to the projection onto the $(\hat{X}_1,\hat{X}_2)$ basis:

$$\begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix} = P \times L \times U \times \begin{bmatrix} M \\ M_d \end{bmatrix}$$  (3.10)

The relation (3.5) and (3.7) are easily translated into matrix terms:

$$P = \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\text{IPD}} \end{bmatrix}; \quad L = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$$  (3.11)

The mixing matrix U is defined by:

$$U = \begin{bmatrix} \cos (\alpha + \beta) & \sin (\alpha + \beta) \\ \cos (-\alpha + \beta) & \sin (-\alpha + \beta) \end{bmatrix}$$  (3.12)

The $\alpha$ coefficient is the one interpreting the ICC and $\beta$ is the corresponding scaling by subband:

$$\alpha[b] = \frac{1}{2} \arccos \left( \text{ICC}[b] \right) \text{ and } \beta[b] = \arctan \left( \frac{c_2 - c_1}{c_2 + c_1 \tan \alpha[b]} \right)$$  (3.13)

This mixing process is a way to add ambience in proportion of the correlation between the two channels of the stereo signal. Using an uncorrelated signal is necessary in order not to degrade the content of the original signal with masking. This step mainly creates a background ambience, yet the issue of localized sound source effect is not tackled.

If we consider that a perfect ICTD translated into a linear IPD (eq 1.8) represents a point source in space (see 4.1 for more details), then in order to enlarge the sound source
localization, that IPD relation has to be degraded somehow. The idea proposed in [12] is to use both the ICLD and the ICC to disrupt that IPD continuity along with adding the uncorrelated signal.

The IPD redistribution over the left and right channels is close to the OPD but better interpreted for ICC values different from 1. It is computed by:

\[
\begin{align*}
\text{CPD1}[b] &= \zeta(c[b] + \text{ICC}[b]e^{j\text{IPD}[b]}) \\
\text{CPD2}[b] &= \zeta\left(\frac{1}{c[b]} + \text{ICC}[b]e^{j\text{IPD}[b]}\right)
\end{align*}
\]

Then applied to the upmix the complete scheme is:

\[
\begin{bmatrix}
\hat{X}_1 \\
\hat{X}_2
\end{bmatrix} = \hat{P} \times L \times U \times \begin{bmatrix} M \\ M_d \end{bmatrix}
\]

with:

\[
\hat{P} = \begin{bmatrix} e^{j\text{CPD1}} & 0 \\ 0 & e^{-j\text{CPD2}} \end{bmatrix}
\]

The geometrical representation of that new phase shift is as follow:

![Diagram](image)

Figure 37: New phase shift CPD1 for sound source enlargement

This method is hierarchical. If the ICLD is important (either \(c[b] \gg 1\) or \(c[b] \ll 1\)), not only the phase remains close to the original IPD but it also aligns with the dominant channel regardless of the ICC value.

In practice if the level difference between the two input channels is large, it is because the sound source is localized either on the left or on the right of the recording device and that the sound source is rather punctual (if the sound source were larger, then the ICLD would not be that extreme). In that case, it is not preferable to try and enlarge the sound source, on the contrary the precise IPD relation is a good approximation and using the dominant channel as a reference enhance the quality. The position of the sound source will be interpreted by the listener thanks to the ICLD.
Figure 38: Large ICLD ($c \gg 1 \Rightarrow \text{CPD}_1 \rightarrow 0, \text{CPD}_2 \rightarrow \text{ICPD}$)

Now if the ICLD is close to 0 then the IPD distribution over the left and right channel is entirely left to the ICC. If the ICC is close to 1 then the IPD is kept and equally distributed over the left and right channels:

Figure 39: Strong coherence between left and right channels (IPD = CPD1 + CPD2)
In a practical sense an ICC close to 1 is representative of a very localized sound source, and as such the IPD must be kept in the upmix process so that the spatial localization of the sound source is well rendered. Moreover, as the level difference between the two channels is low, it means that the sound source is in front of the listener. In that case, as the ILD is no more reliable enough to position the sound source, the ICC is used to ensure that function through the IPD.

The last case is the one were the ICLD is close to 0 and the ICC is low. Then the IPD is reduced:

![Diagram](image.png)

Figure 40: Low coherence between left and right channels (IPD > CPD1 + CPD2)

That situation is the most interesting one. The low ICC value is the definition of a diluted sound source which in practice corresponds to a ambience noise rather than an actual sound source. As the low ICC value pulls the CPDs close to 0, and keeping in mind that the ICLD is also close to 0, this leads to the synthesis of a perfect background noise equally distributed in loudness and synchronized between the two output channels.

As this upmixing process is theoretically promising, it will be the one used for the EVS stereo extension codec. Different ways to use it in agreement with the bit cost quotas are presented in the next subsection.

### 3.4.2 Ambience through reverberation

The use of ambiance synthesis in parametric stereo coding helps creating a better audio rendering. As we usually are listening to music or conversations captured in rooms, recreating the impact of that room on the sound itself makes it more natural for us to hear. This idea is the one developed in [11].

The main idea is to use a decorrelation filter such as in 3.4.1 but also to enhance the effects of such an ambiance synthesis by adding delays through comb filters representing room reverberations. As the algorithm in [11] is based on QMF, those delays are implemented as fractional delays on each QMF subband.

However this technique, implemented in QMF domain, has not been tested for the EVS stereo extension.
4 EVS stereo extension at 32 kbit/s

The EVS codec developed in 3GPP is a codec used to reduce the bit rate with an optimized quality [10]. This codec is designed for monaural signals at given sampling frequencies and the bit rate reduction range goes from 5.9 kbit/s to 128 kbit/s. The sampling rate falls into 4 categories: narrow band (NB) at 8 kHz, wide band (WB) at 16 kHz, super wide band (SWB) at 32 kHz and full band (FB) at 48 kHz. Our case study was WB signals. For such applications the desired bit rate is often set to 24.4 kbit/s as that is the most used for telecommunications, thus leaving [5.9, 7.2, 8.0, 9.6, 13.2, 16.4, 24.4] kbit/s available. Yet for this study we choose a 32.0 kbit/s in order to evaluate the extension with respect to the dual mono codec. An uncompressed WB monaural signal encoded in 16 bit PCM has a native bit rate of 256 kbit/s, thus for stereo the aim is reducing the bit rate from 512 kbit/s to 32 bit/s, so reaching a compression factor of 1/16 with the least audible degradation.

4.1 EVS dual mono

As the codec is designed for monaural signals, stereo encoding can be managed as dual mono: left and right channels are compressed simultaneously:

![Figure 41: EVS codec used as dual mono encoder](image)

This method is regardless of the channels content so the bit rate allocation has to be equally divided into the encoding of both channels (i.e. $R_1 = R_2$). Thus a 16.4 kbit/s rate is used on both channels at the encoder (EVSe) and then transmitted to the receiving end (EVSD) where the two compressed signals are decoded and used as left and right channels for the stereo output. This model is simple yet it yields good results for non correlated signals. As no real synthesis algorithm is needed, the stereo effects are mainly kept through the compression embedded in each channel, the output has a clean spatial rendering as long as the bit rate is high enough. The main issue is the quality of the encoding decoding process, but for 16 kHz sampling rate, the majority of the stereo signals are well transmitted without audible quality loss.

In order to enhance the quality of the transmitted stereo file, the idea is to use a parametric coding approach and evaluate the pros and cons of transmitting a downmix and inter channel cues. As only one "sound" file will have to be compressed, the bit rate used can be less restrictive, but the inter channel cues have to fit in the complementary bit rate. In order to reach a sufficient quality, it is essential for the downmix signal to be of great quality itself otherwise the upmix is bound to have poor quality. The work on the new hybrid+ downmix was part of that requirement. This new parametric model is the stereo extension draft of the EVS codec.
4.2 EVS stereo extension

The parametric model for the stereo extension follows the parametric coding outline described at the beginning of this report. The main architecture of that model is represented below:

![Figure 42: EVS stereo extension](image)

The content is mainly in the downmix (DMX) whereas the spatial information are transmitted as the cues. Two main approaches have been chosen, one is to focus more on the quality of the content and the other to equally distribute the bit allocation to the spatial cues and the content. The first approach will be referred to as the asymmetrical EVS stereo extension (aEVSse) and the second the symmetrical EVS stereo extension (sEVSse), they are presented and evaluated in the following subsections.

4.2.1 Asymmetrical EVS stereo extension

In order to enhance the quality of the downmix transmission by reducing its compression, the EVS encoder is used at $R_1 = 24.4$ kbit/s, leaving $R_2 \approx 8.0$ kbit/s for the spatial cues. As depicted in subsection 3.2, several options are available to quantify the cues, yet as strong quantification will occur to meet the desired rate, we have to ensure that the quality of the spatial image is sufficient. This subsection describes one possible process that showed good results.

The IPD is encoded using the 4-bit codebook for the first 20 subbands for each frame (a 1-bit value is sufficient for the first subband where IPD values are either 0 or $\pi$). The ICLD is encoded with the 5-bit codebook for the first subband, then the 4-bit codebook for the subband from 2 to 20 and then the 3-bit codebook for the rest. The ICC is encoded with the 3-bit codebook from the 6th subband (below its value is always 1 or close to 1 for the 4th subband, so ICC = 1 is used for the upmix) to the 20th subband. The ICLD and ICC are transmitted every two frames and values are kept constant for each frame to the next not transmitted. Thus the overall cost is $R_2 = 8.1$ kbit/s. A basic upmix (3.8) is used for the first 5th subband and the last 15th, a complete upmix (3.16) for the central subband.

<table>
<thead>
<tr>
<th>Subband number</th>
<th>1</th>
<th>2-5</th>
<th>6-20</th>
<th>21-35</th>
<th>$R_2$ (bit/f, kbit/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPD</td>
<td>1-bit</td>
<td>4-bit</td>
<td>4-bit</td>
<td>NT</td>
<td></td>
</tr>
<tr>
<td>ICLD</td>
<td>5-bit</td>
<td>4-bit</td>
<td>4-bit</td>
<td>3-bit</td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>NT</td>
<td>NT</td>
<td>3-bit</td>
<td>NT</td>
<td></td>
</tr>
<tr>
<td>Total per frame (bit/f)</td>
<td>3.5</td>
<td>24</td>
<td>112.5</td>
<td>22.5</td>
<td>162.5, 8.1</td>
</tr>
</tbody>
</table>

![Figure 43: Bit allocation for the aEVSse (NT stands for not transmitted)](image)
4.2.2 Symmetrical EVS stereo extension

For WB signals the EVS codec can be sufficient at $R_1 = 16.4 \text{ kbit/s}$ for transmitting a good quality downmix, thus the symmetrical extension was adapted to this rate and twice as more cues could be transmitted in order to improve the spatial image. For a bit rate of 24.4 kbit/s this approach would however become less "symmetrical" with bit rates such as $R_1 = 13.2 \text{ kbit/s}$ and $R_2 = 9 \text{ kbit/s}$.

The IPD is encoded using the 4-bit codebook for the first 25 subbands for each frame, which enhances the spatial rendering for complex signal (see subsection 3.2). The ICLD is encoded with the 5-bit codebook for the first subband, then the 4-bit codebook for the subband from 2 to 25 and then the 3-bit codebook for the rest. The ICC is encoded with the 3-bit codebook from the 6th subband to the 35th subband. All the cues are transmitted for each frame. Thus the overall cost is $R_2 = 15.9 \text{ kbit/s}$. A basic upmix (3.8) is used for the first 5th subband, a complete upmix (3.16) for the rest of the subband, and as the IPD is kept at 0 for the last 10th subband $P = I_2$.

<table>
<thead>
<tr>
<th>Subband number</th>
<th>1</th>
<th>2-5</th>
<th>6-25</th>
<th>26-35</th>
<th>$R_2$ (bit/frame, kbit/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPD</td>
<td>1-bit</td>
<td>4-bit</td>
<td>4-bit</td>
<td>NT</td>
<td></td>
</tr>
<tr>
<td>ICLD</td>
<td>5-bit</td>
<td>4-bit</td>
<td>3-bit</td>
<td>3-bit</td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>NT</td>
<td>NT</td>
<td>3-bit</td>
<td>3-bit</td>
<td></td>
</tr>
<tr>
<td>Total per frame (bit/frame)</td>
<td>6</td>
<td>32</td>
<td>220</td>
<td>60</td>
<td>318, 15.9</td>
</tr>
</tbody>
</table>

Figure 44: Bit allocation for the sEVSse

Due to the higher transmission rate, the quality of the downmix decoded in the aEVSse is better than the one in the sEVSse, yet as the ICLD is only transmitted one every two frames, it sometimes degrades the overall quality of the aEVSse and the sEVSse yields better results. For both methods the spatial image reconstruction is not sufficient for complex signals such as orchestra and the sound is still heard at the center of the head. The decorrelated signal creates an effect of ambiance noise that enlarges the sound image, yet it often happens that the level of that signal is too high and becomes annoying, usually when the downmix signal has a low level. In order to reduce that effect a threshold as been applied cancelling that signal when the downmix signal is too weak. The threshold is seen as a minimum square root of the energy of the downmix per frame: $E_{th} = 321$, in average all the frequency bins components shall be greater than 1.

$$E_f = \frac{1}{321} \sqrt{\sum_{k=1}^{321} |M[k]|^2}$$  \hspace{1cm} (4.1)

The two extensions, symmetrical and asymmetrical, are both interesting and switching from one to the other might prove to be the best solution after using a criteria such as the correlation between both channels. The aEVSse can be easily translated into a 24.4 kbit/s by switching the coding rate of the downmix to 16.4 kbit/s, whereas the sEVSse would need some adjustments. As the EVS codec has a 12 ms delay and that the analysis frame length is 40 ms this adds a 20 ms delay due to the first overlap, plus the 8 ms to fall back on a 20ms multiple. Thus the overall delay of the the two methods should be 40 ms and meets the requirement for stereo which is fixed at 50 ms.
5 Spatial coding for stereo application

Using virtual sound sources is an interesting approach in order to reduce even more the bit rate. Interpreting the sound as a virtual sound source and transmitting only the position of that source is an efficient way to encode stereo effects for simple signals such as single talkers. The two following sections are applications of that idea. Different and more detailed techniques can be found in [19] [21], also an international workshop has led to DirAC (Directional Audio Coding) which is a technique for spatial audio processing using similar ideas [22],

5.1 Source localization

For the spatialization of sound, estimating a virtual sound source position from the input signals can be useful for parametric coding. If we consider the 2D horizontal plane surrounding a listener, the sound source position can be given by two polar coordinates \((r, \theta)\):

\[
\text{ITD}(\theta) = \frac{a \sin \theta}{c} \quad (5.1)
\]

This frequency independent law is known to yield good results in terms of spatial localization. Furthermore the ILD can also be approximated using polar coordinates [9] :

\[
\text{ILD}(\theta, r) = \frac{80\pi fr \sin \theta}{c \ln 10} \quad (5.2)
\]

Using an ITD estimator, see subsection below, a virtual sound source position located with \(\theta\) can be defined and using only that angle clues can be synthesized at the decoder’s end :

\[
\text{ITD}(\theta) = \text{ITD}_{\text{max}} \cdot \sin \theta \\
\text{ILD}(\theta, r) = \text{ILD}_{\text{max}} \cdot g(r) \sin \theta \quad (5.3)
\]

The value assigned to \(\text{ITD}_{\text{max}}\) is fixed at 0.630 ms which is the limit for perceptual separation of two impulses, i.e. for larger ITD the listener will hear two distinctive sounds.
and won’t be able to interpret the sound as coming from one single sound source. The ILD approximation is more controversial because the coloration of sound due to HRTFs mainly prevails for sound source localization. As opposed to the real ITD which is exactly defined and can be used in the digital domain as a delay between the two channels, the ILD cannot be absolutely defined for the level difference in the digital domain is bound to the quantization. Thus a relative ILD computed from the signals themselves can be used to normalize the approximation, erasing at the same time the distance dependence \(g(r)\). Using the properties of the input signals to defined a reference ILD (ILD\(_{\text{max}}\)) can be effective:

\[
\text{ILD}(\theta) = \text{ILD}_{\text{max}} \cdot \sin \theta
\]  

(5.4)

The estimation of the value assigned to ILD\(_{\text{max}}\) is developed in subsection 4.3.

This approach is only valid for single sound sources and artifacts will occur when two separate sound sources are present in the input signals. Yet the advantage is that only the angle \(\theta\) is transmitted and takes only one value per frame, thus reducing bitrate.
5.2 ITD estimation

ITD is a strong cue for sound source localization for the low frequency region (f < 2 kHz). The most classic and robust estimator is to look for the maximum of the crosscorrelation between the left and right channel in the time domain:

\[
\Phi(x_1, x_2, \tau) = \sum_{n=0}^{N-\tau-1} (x_1[n + \tau] \cdot x_2[n])
\]  
(5.5)

The crosscorrelation representing the cross product of the two channels, one being delayed consecutively by \(\tau\) samples, its maximum is reached when the two channels are superposed, in other words when the delay \(\tau\) is equal to the corresponding ITD. As our ITD takes its value between -630 ms and 630 ms, the search for a maximum can be reduced to corresponding sampled values using the sampling frequency \(f_s\):

\[
\tau \in [-d; d]
\]

where \(d = 6.3 \times 10^{-4} f_s\). The sampled ITD (ITD\(_{s,f}\)) for a specific frame is consequently evaluated by:

\[
\text{ITD}_{s,f}(x_1,f, x_2,f) = \max_{-d\leq \tau \leq d} \sum_{n=0}^{N-\tau-1} (x_1[n + \tau] \cdot x_2[n])
\]  
(5.6)

Figure 46: ITD estimation from the crosscorrelation, \(f_s = 44,1\) kHz, \(d = 28\) samples.

This ITD estimation remains noisy, so a pre and post processing has to be done to smoothen the results. First low energy portions of the signals should not be taken into account to avoid computing ITD on empty (noise dominant) portions of the signals. One way is to evaluate the average energy of the signal over time and use a threshold such that if the energy inside a frame is lower than that threshold, no ITD is estimated and the first previous non empty frame is used as a default value. Second even though non representative values have been neglected, the estimation is still noisy. A two step smoothing process is then used to average a continuous ITD. A first 10th order moment over a small number of frame \((2l_1 + 1)\) is computed 3 times, and then a 10th order moment is computed one time over a large number of frame \((2l_2 + 1, l_2 >> l_1)\) to ensure some continuity and avoid artifacts. This leads to the following estimations and approximations:
\[\text{ITD}_{s,f,\text{smth1}} = \left( \frac{1}{2l_1 + 1} \sum_{\eta=f-l_1}^{f+l_1} \text{ITD}_{s,\eta}^{10} \right)^{1/10}\]

\[\text{ITD}_{s,f,\text{smth2}} = \left( \frac{1}{2l_2 + 1} \sum_{\eta=f-l_2}^{f+l_2} \text{ITD}_{s,\eta,\text{smth1}}^{10} \right)^{1/10}\]

(5.7)

Figure 47: ITD smoothing

As we can see from figure 18, the smoothing process reduces the maximal values taken by the ITD, here from 28 to 25, so a small adjustment has to be made afterwards. Furthermore the ITD sensitivity in terms of sound source localization is rough for lateral positions where the localization accuracy is 15 degrees [25] so the adjustment doesn’t require too much precision. However the accuracy of localization in front of the listener is 1 degree [25] so error around that position (ITD = 0) has to be limited to avoid confusion.
5.3 ILD estimation

The geometrical approximation of the ILD is not well founded due to the filtering process depicted in section 4. Yet if we compute the total energy of the filters acting on the sound source when reaching our ears (HRTF), a sin law doesn’t seem to be far fetched:

![Figure 48: Total Energy measured for HRTFs evolving around the head](image)

Even though the symmetry between the half space facing the ear (azimuth $\in [0, 180]$) and the half space hidden by the head (azimuth $\in [180, 360]$) is not quite respected, a sin law could be used to roughly synthesize a basic HRTF effect. Then as both energies received by left and right ears will evolve in a sin way, the difference should also be seen as a sin law. Using the definition (1.1) to compute the ILD between signals received by left and right ear, the simple sin law as given in (5.3) is proven relevant:

![Figure 49: Global ILD computation over all frequencies for binaural signals](image)
Depending on the distance between the sound source and the listener head, the maximum value reached by the ILD is bound to evolve. Also a general ILD is relative to the quantification and the level of the signal, a reference ILD taken as the maximum of the total ILD over the frames could be used in case of non real time applications.

\[ \text{ILD}_{\text{ref}} = \max_f \left( \sum_b \text{ILD}_f[b] \right) \]  

(5.8)

Using (4.2) we can arbitrarily rescale it to better fit the signal:

\[ \text{ILD}_{\text{max}} = \frac{80\pi}{c \ln 10} \text{ILD}_{\text{ref}} \]  

(5.9)

This leads to the following modeling:

\[ \text{ILD}_f = \text{ILD}_{\text{max}} \cdot \sin \theta(f) \]  

(5.10)

And using (3.7) notations, the upmix process is defined as:

\[
\hat{X}_{1,f}[k] = c_{1,f}(\theta_f)M_f[k] \\
\hat{X}_{2,f}[k] = c_{2,f}(\theta_f)M_f[k] \cdot e^{-2\pi k f_s \text{ITD}(\theta_f)/N_{fft}}
\]  

(5.11)

Figure 50: ILD modeling using \( \text{ILD}_{\text{max}} \) and our angle \( \theta \)

In the first portion of the signal, only the spikes in ILD represent the actual person talking on the track, even though the smoothing process reduces the actual measured ILD the synthesis is quite good to listen to. For real time application one could use tables with predefined reference values for the ILD or subband ILD tables given for chosen angles.

That spatial method for parametric coding could be used for telephonic conversation transmission. A simple stereo upmix such as described above would render a half realistic auditory scene for the listener, and thus might be less tiresome for a long conversation. Due to the chronology of the internship no ICC effect had been seen at that point, so adding them to recreate some background noise non localized could enhance the quality of the decoding process.
6 Spatial coding for binaural application

6.1 Virtual 3D sound

Our ability to interpret sound source positions in space when listening is the basis of 3D sound. Before reaching our ears, the incoming sound is transformed and colored in a very specific way that our brain has learned to identify with a position in space. Those transformations in the frequency domain are called Head Related Transfer Functions (HRTFs) or Head Related Impulse Response in the time domain (HRIRs). By measuring those functions we can recreate a 3D image of the sound for any position of the sound source in space. This technique is referred to as binaural sound when the 3D image is perceived using headphones.

HRTFs are rich filters that actually hold the spatial ILD and ITD cues. If the measurement is made on a sphere around the test subject, then the ITD is embedded in the HRTF. The coloration difference between the left and right ear HRTFs for one location of the sound source is the real ILD. Often measurements are made on a sphere around subject and so a 3D image synthesis at a specific distance is easily achievable. With the angular position of the sound source \((\theta, \phi)\), filtering a sound using the corresponding HRTF\(_1\)\(_2\)\((\theta, \phi)\) and sending it to the corresponding ear will suffice.
6.2 From Stereo to 3D

As depicted in section 3, for single virtual sound sources we can approximate their positions in space using the ITD transposed into an angle $\theta$. Now using that same angle, we can project the sound in 3D. HRTFs are usually measured at a given angle in space around the head of the subject: $\text{HRTF}_d(\theta_i, \phi_i)$, where the ear is approximately considered at the center of a sphere. In order to be able to create a real continuous 3D image we need to interpolate those discrete HRTFs into continuous ones: $\text{HRTF}_c(\theta, \phi)$.

This can be done using the ambisonic theory (a more detailed model can be found in [20]). Ambisonic relies on a 3D representation of the sound field using a 3D Bessel-Fourier decomposition on a basis called spherical harmonics. The basis is given by:

$$\hat{Y}_m^n(r, \theta, \phi) = r^n Y_m^n(\theta, \phi)$$

where $n \in [0, N]$ and $m \in [-n, n]$.

In those expressions, $P_m^n$ are the associated Legendre polynomials defined as:

$$P_m^n(x) = \begin{cases} 
(-1)^m \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}} P_m^n(\sin \phi) \cos (m\theta) & \text{for } m < 0 \\
\sqrt{\frac{(2n+1)}{4\pi}} P_0^n(\sin \phi) & \text{for } m = 0 \\
(-1)^m \sqrt{\frac{(2n+1)(n+m)!}{4\pi(n+m)!}} P_m^n(\sin \phi) \sin (|m|\theta) & \text{for } m > 0
\end{cases}$$

Those polynomials are neither normalized nor semi-normalized but used as defined in (4.3).
Using our basis we can project our HRTF\(_d\) on the corresponding angles and extract spherical HRTFs:

\[
\text{HRTF}^{n,m}_{\text{spher}} = \sum_{\theta_i, \phi_i} \hat{a}(\theta, \phi) \text{HRTF}_d(\theta, \phi) \cdot Y^m_n(\theta, \phi) \quad (6.4)
\]

The weighting coefficients \(\hat{a}(\theta, \phi)\) are approximated from the real integral where the actual coefficients \(a(\theta, \phi)\) are given by:

\[
a(\theta, \phi) = \sin(\phi) d\theta d\phi \quad (6.5)
\]

Leading to values for \(d\theta = d\phi = 15^\circ = 0.2618\) rad to

\[
\hat{a}(\theta_i, \phi_i) = \frac{(0.2618)^2}{3} \sin(\phi_i) \quad (6.6)
\]

Using this new HRTF basis we can extract an interpolated HRTF at any given angle:

\[
\text{HRTF}_{\text{extract}}(\theta, \phi) = \sum_{n,m} \text{HRTF}^{n,m}_{\text{spher}} \cdot Y^m_n(\theta, \phi) \quad (6.7)
\]

As the HRTFs are defined for one ear, we need to interpolate two HRTFs for the left (HRTF\(_1,\text{extract}\)) and the right ear (HRTF\(_2,\text{extract}\)). Yet to simplify the algorithm, the head is supposed to be symmetrical and so the HRTFs too for the two ears, which leads to the following relation (considering that both ears are situated simultaneously at the center of the decomposition sphere, which is roughly true):

\[
\text{HRTF}_{2,\text{extract}}(\theta, \phi) = \text{HRTF}_{1,\text{extract}}(-\theta, \phi) \quad (6.8)
\]

Thus in our 2D system, using the angle \(\theta\) we can project the initial stereo input into a 2D sound. One has to be careful that HRTFs derived from HRIRs might already include ITD. For instance for the IRCAM database, the HRIR are equalized but not zero phased. In order to simplify the extrapolation problem in terms of mathematical complexity, it is better to separate on the one hand zero phased HRIRs and interpolate them as depicted above, and on the other hand to reimplement the ITD on the filtered signals afterwards. This leads to the following up mix for a given frame \(f\), a given bin \(k\) and angle \(\theta_f\):
\[ \hat{X}_{1,f}[k] = M_f[k] \cdot \text{HRTF}_{1,\text{extract}}(\theta_f)[k] \]
\[ \hat{X}_{2,f}[k] = M_f[k] \cdot \text{HRTF}_{1,\text{extract}}(-\theta_f)[k] \cdot e^{-2\pi k f_s \text{ITD}(\theta_f)/N_{\text{fft}}} \]

As the idea presented in section 4, this process could be applied to telephonic audio transmission. The complexity at the decoding end would require performant cellphones but the rendering would be a lot more realistic than the one synthesized in the previous section. Another barrier is the authenticity of the HRTF filters, which is a constant problem in applications such as those. HRTFs are unique and using general filters can lead to audio artefacts if they don’t match the ones of the listener. Studies on individualizing HRTFs using subband equalization or known corporal related factors could help improve such technologies. Similar ideas are developed in [17].
7 Conclusion

The new downmixing technique developed in the first part of this thesis is an important improvement comparing with preexisting methods for its robustness, to be understood as uniformly good quality for many types of stereo signals. For applications such as compatibility from stereo to mono which are required in certain systems, that downmix can be used to ensure a good conversion from the original stereo file to the mono system, which is the same issue as for very low bit rate transmission of stereo where only one sum signal can be encoded. The idea behind the hybrid+ downmix could also be extrapolated to other downmixing problems such as multichannel conversion to stereo.

The EVS stereo extensions that have been developed during this thesis are promising and could be used to fully develop a complete stereo solution. Finding the right way to switch from the two methods for wide band signals could be the simplest solutions to a robust codec. It is probable that for super wide band signals, the asymmetrical EVS stereo extension will lead to better results because the transmission rate of the downmix has a greater impact on those signals and other rates than 32 kbit/s should be considered. For 24.4 kbit/s the asymmetrical EVS stereo extension is the best choice. Techniques described in the last parts of the report could be added to the EVS extension either to simplify the cue transmission in case of simple signals such as single speakers, or create binaural rendering for specific applications.
8 Appendix A: Sound sample treatment before test

The goal of this appendix is twofold. Firstly we present how the sound files samples have been treated so that the subjective tests are unbiased and also the test itself less tiresome. Secondly such a description could be used to help future tests to be prepared.

Once the basis of sound samples is chosen (roughly 8 to 10 seconds of sound per file), all the files are normalized in terms of audio levels and formats:

- To prevent any issue due to the .wav format signature, all the files are converted into raw PCM files (CopyAudio batch). Then they are leveled at -26 dB overload using the maximum absolute value of the stereo signal (maximum over the left and right channel) (stereoop and sv56demo batch). Such a normalization will avoid manual level changes during the test.

- Fade in and fade out are applied to start and ending of the files if needed (CoolEdit). This is mainly to avoid strong clips or cut during the listening process, which makes the test more pleasant.

- All the reference samples are then kept and will be used as reference if needed in the subjective test.

- Reference samples go through filtering process and output files are saved as raw PCM files (Matlab and/or C).

- All filtered signals and references are encoded back in .wav (the STEP program used for the tests needs .wav files as input) (CopyAudio batch).

- For our case of one monaural signal, the output was duplicated and thus a dual mono signal was used for the test with diatonic listening (stereoop and CopyAudio batch).
References


