Parametric representation of multichannel audio based on Principal Component Analysis

Manuel Briand, David Virette, Nadine Martin

Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Parametric Representation of Multichannel Audio Based on Principal Component Analysis

Manuel Briand[1], David Virette[1], and Nadine Martin[2]

[1] France Telecom R&D, Speech and Sound Technologies and Processing Lab – 2, avenue Pierre Marzin
22307 Lannion Cedex, France
{manuel.briand,david.virette}@francetelecom.com

[2] Laboratoire des Images et des Signaux (CNRS UMR 5083) – 961, rue de la Houille Blanche BP 46
38402 S[t] Martin d'Hères Cedex, France
nadine.martin@lis.inpg.fr

ABSTRACT

Low bit rate parametric audio coding for multichannel audio is mainly based on Binaural Cue Coding (BCC). In this paper we show that the Unified Domain Representation of multichannel audio, recently introduced, is equivalent to BCC scheme in stereo coding context. We also discuss another method, called multichannel audio upmix, which classically converts existing two-channel stereo to five-channel audio. More precisely, we focus on existing PCA-based upmix method. Starting from PCA approach, we propose a general model that may be applied both to parametric representation of multichannel audio signals and upmix methods. Moreover, we apply the analysis results to propose a new parametric audio coding method based on frequency subbands PCA processing.

1. INTRODUCTION

With the introduction of multichannel audio playback systems for consumer use with 5.1 playback system or above, multichannel audio content has become necessary for low data rate application such as HD-Radio, Audio on Demand, etc. Therefore, two close audio processing methods are currently considered to deliver multichannel audio for low bit rate applications.

The first method is denoted as multichannel audio coding. Matrix surround coding schemes and parametric audio coding schemes are the two main multichannel audio coding techniques currently used. Matrix surround coding scheme such as Dolby Pro Logic [1] consists in matrixing the channels of the original multichannel signal in order to reduce the number of signals to be transmitted. Passive or active decoding could be achieved to build a multichannel signal

perceptually as close as possible to the original multichannel signal. Nevertheless, this multichannel audio coding method cannot deliver multichannel audio with a data rate acceptable for most networks. That is made possible with low bit rate parametric audio coding mainly based on Binaural Cue Coding [2] (BCC). This coding scheme represents multichannel audio signals by one or several downmixed audio channels plus spatial cues extracted from the original channels. Recently, the Unified Domain (UD) representation of multichannel audio has been introduced [5]. This lossless and invertible transformation (UDT) is equivalent to a rotation in a multidimensional complex space with basis defined by the frequency domain channels of the multichannel signal. In this paper we show how the UDT could be equivalent to BCC scheme in a stereo coding context.

A second multichannel audio processing method, called upmix, classically converts the existing stereo audio contents into five-channel audio. The spatial characteristics and the coherence of the stereo signal are used to synthesize a multichannel audio signal compatible with home cinema setup. More precisely, we focus on existing PCA-based upmix method [6]-[7]. The rear channels are considered as ambience channels defined as diffuse surround sounds and the center front channel corresponds to the sources panned across the original stereo channels. The first step of the upmix algorithm in [6] consists in a Principal Component Analysis (PCA) of the stereo signal. The PCA is equivalent to a rotation of the stereo signal coordinate system and results in one principal component signal and a remaining signal. The principal component signal corresponds to the dominant source present in the original stereo. Then, the center channel results from the weighting of this principal component by a coefficient derived from the rotation angle of the coordinate system. The rear channels result from the weighting of the remaining signal by a coefficient derived from the correlation coefficient of the stereo signal. A time-domain subband processing of this upmix method has recently been proposed in [7].

Starting from PCA approach, we propose a general model that may be applied both to parametric representation of multichannel audio signals and upmix methods. Moreover, we apply the analysis results to propose a new parametric audio coding method based on frequency subbands PCA processing. This paper is organized as follows. In section 2 the equivalence of UDT to BCC scheme in stereo coding context is underlined. A general model of multichannel audio signals is exposed in section 3 to outline the repartition of stereo signals eigenvalues and then derive a parametric representation of multichannel audio based on PCA. Finally, a new parametric audio coding method is exposed in section 4.

2. BCC AND UDT EQUIVALENCE

The equivalence of UDT and BCC scheme is considered here in a stereo signals coding context. The BCC approach [2] is based on the extraction and coding of auditory localization cues. Actually, during the auditory localization process, the signals arriving at the eardrums present inter-aural time and level differences (ITD and ILD) also called binaural cues. The human auditory system could be then considered as a time-frequency analyzer. The spectral resolution of the auditory system can be described by a filter bank with filter bandwidths that follow the ERB (Equivalent Rectangular Bandwidth) scale considered as an approximation of the critical bands [3]. Under these auditory perception considerations, binaural cue coding scheme realizes the extraction and coding of spatial parameters and perceptual audio coding of the downmix mono signal (in case of stereo input). The downmix signal is generated by adding the input channels in the subband domain and multiplying the sum with a factor in order to preserve signal power (see [2] for more details).

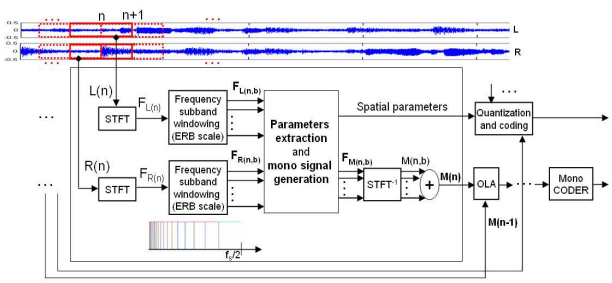

Figure 1: Binaural Cue Coding scheme applied to stereo.

Inter-channel time and level differences (ICLD and ICTD) are extracted from each analyzed signal block and each frequency subband. ICTD cue could also be considered as phase difference (ICPD) between the analyzed channels as J. Breebaart and al. have suggested in [4]. Moreover, inter-channel coherence (ICC) is also extracted in order to measure the spatial diffuseness of the sound sources. Considering the Short Time Fourier Transform (STFT) of stereo channels blocks n, the frequency-domain signals $F_L(n)$ and $F_R(n)$ are divided

into nonoverlapping subbands by frequency windowing (bins grouping) according to ERB scale (cf. Figure 1). The spatial cues expressions can be written as:

$$\begin{cases} ICLD(n,b) = 20\log_{10}\left(\dfrac{\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{R(n)}(k)\right|^2}{\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{L(n)}(k)\right|^2}\right) \\ ICPD(n,b) = \arg\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{L(n)}(k)\cdot F^*_{R(n)}(k)\right) \\ ICC(n,b) = \dfrac{\Re\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{L(n)}(k)\cdot F^*_{R(n)}(k)\right)}{\sqrt{\left(\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{L(n)}(k)\right|^2\right)\left(\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{R(n)}(k)\right|^2\right)}} \end{cases} \quad (1)$$

where * denotes the complex conjugation. The ICC parameter is defined as the normalized cross-correlation coefficient, derived from the cross-spectrum of the stereo channels, after phase alignment of the analyzed stereo blocks according to the ICPD.

Another multichannel audio processing method has been introduced by K.M. Short and al. in [5] which propose a unified domain representation of multichannel audio signals. As for low-complexity BCC scheme, the first step of the transformation to the Unified Domain consists in the transformation of each channel to the frequency domain with Discrete Fourier Transform. Then, a complex rotation of the left and right spectra $F_C(k) = \left|F_C(k)\right| e^{j\phi_C(k)}$ with channel C = L or R is performed for each frequency bin k:

$$\begin{bmatrix} e^{-j\phi_L(k)}\cos\sigma(k) & e^{-j\phi_R(k)}\sin\sigma(k) \\ -e^{j\phi_R(k)}\sin\sigma(k) & e^{j\phi_L(k)}\cos\sigma(k) \end{bmatrix}\begin{bmatrix} F_L(k) \\ F_R(k) \end{bmatrix} = \begin{bmatrix} M(k) \\ 0 \end{bmatrix} \quad (2)$$

The Unified Magnitude and the rotation angle are defined according to:

$$M(k) = \sqrt{\left|F_L(k)\right|^2 + \left|F_R(k)\right|^2} \quad (3)$$

$$\begin{cases} \cos\sigma(k) = \dfrac{\left|F_L(k)\right|}{M(k)} \\ \sin\sigma(k) = \dfrac{\left|F_R(k)\right|}{M(k)} \end{cases} \quad (4)$$

Benefit of UDT is that it unifies components of a source signal that may be distributed over many channels. The UDT rotation angle could be seen as a spatial parameter equivalent to the ICLD cue of BCC extracted for each frequency bins of the stereo channels spectra. Indeed, we can derive from equation (4):

$$\tan\sigma(k) = \frac{\left|F_R(k)\right|}{\left|F_L(k)\right|} \quad (5)$$

So we can define a relation between the rotation angle of the UD transformation and the ICLD cue of BCC as:

$$\tan\sigma(k) = 10^{\frac{ICLD}{10}} \quad (6)$$

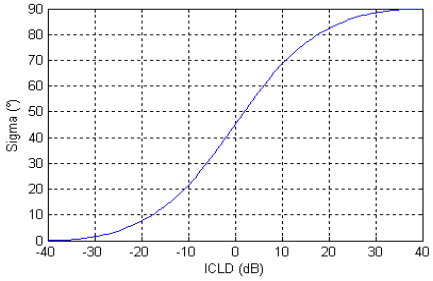

Figure 2: Unified Domain rotation angle (in degrees) function of the ICLD cue (in decibels).

The UDT could be seen as an equivalent of the BCC scheme in a stereo coding context. Actually, the Unified Magnitude $M(k)$ could be considered as a downmix signal of the stereo input. Then, this encoded signal by a traditional mono coder could be transmitted plus rotation angle $\sigma(k)$ in order to distribute the magnitude over the decoded channels. Moreover, inter-channel phase difference ($\phi_L(k)$ - $\phi_R(k)$) of the left and right channels could be also transmitted as ICPD for BCC scheme.

In this section, we show how the rotation angle used in the UDT could be seen as a function of the ICLD cue used in BCC scheme. Moreover, the rotation angles use in the UDT correspond to the spatial positions of the spectral components while magnitude is related to the amplitude of these components [5]. This approach is very close to Principal Component Analysis (PCA) which is the main process of upmix method in [6]-[7].

Based on PCA approach, next section presents a general model for multichannel audio signals which results in a parametric representation of multichannel audio.

3. PARAMETRIC REPRESENTATION OF MULTICHANNEL AUDIO BASED ON PCA

Multichannel audio signals could be considered as studio (artificial) or live (natural) recorded signals [8]. Live recording involves many different setups and microphones types which determine the amount of interferences and reverberation on each channel. Moreover, because real spaces are decorrelated, decorrelated signals yield a sense of realistic ambience. Ambience is complex audio content, perceptually background and very heterogeneous that could be defined as "the sound of the place in which sources are". Such audio content includes acoustic effects of reverberant volumes and reflective features plus the background i.e. the acoustic accumulation of many small sources that are not the identified sources of interest; for example, audience noise. In studio recording, sound sources (instruments) are individually recorded and then processed. The processing of recorded sound sources consists in applying panning functions to the sound sources and then mixing them with synthetic reverberation. In order to increase the perception of spaciousness, weakly correlated reverberation impulse responses are used. Under these assumptions, we define a general model for multichannel audio signals.

3.1. Multichannel audio model defined as directional sources and ambiences

The multichannel audio model is defined according to next suggestions. We assume the presence of *D* directional sources, easily localisable, panned - distributed - into *M* channels. Moreover, we consider *M* ambience signals i.e. one by channel, which are weakly correlated with directional sources. The model of each channel is defined as the sum of direct sources, weighted according to their spatial perceived positions, and one ambience signal. Signal channels following this model are strongly correlated with the presence of directional sources among several channels whereas ambience signals, defined as diffuse sounds, are decorrelated from one channel to another. So, the time domain multichannel signal $\underline{S}_M = \left(S_1, \cdots, S_M\right)$ can be written as:

$$S_m(t) = \sum_{d=1}^{D} \left[g_{m,d}(t) \cdot S_d(t) \right] + A_m(t) \qquad (7)$$

where: $m \in [1, \cdots, M]$ and $g_{m,d}(t)$ are the panning functions (gains) applied to the directional sources S_d of the m[th] channel. By using equation (7) and the vector/ matricial notations: $\underline{S}_D = (S_1 \ \cdots \ S_D)$, $\underline{A} = (A_1 \ \cdots \ A_M)$ and $G = \begin{pmatrix} g_{11} & \cdots & g_{1D} \\ \vdots & \ddots & \vdots \\ g_{M1} & \cdots & g_{MD} \end{pmatrix}$, we can derive the covariance matrix of such a centred (^{C}) multichannel signal:

$$\begin{aligned} R_{S_M} &= \mathrm{E}\left[\underline{S}_M^C \cdot \underline{S}_M^{C\,T} \right] \\ R_{S_M} &= G \cdot \mathrm{E}\left[\underline{S}_D^C \cdot \underline{S}_D^{C\,T} \right] \cdot G^T + \mathrm{E}\left[\underline{A}^C \cdot \underline{A}^{C\,T} \right] \\ R_{S_M} &= G \cdot R_{S_D} + R_A \end{aligned} \qquad (8)$$

where ^{T} denotes the matricial transposition. So, the multichannel signal covariance matrix R_{S_M} is function of the directional sources covariance matrix R_{S_D} and the ambience signals covariance matrix R_A.

Described model assumptions differ from traditional blind source separation (BSS) model (model expression in [9]) in which sources are convolved with FIR filters that model the room transfer function between the m[th] sensor and the d[th] source. Actually, BSS wishes to estimate dry sources from sources mix. Subspace methods, such as describe in [10], whish to reduce incoherent reflections in rooms or more generally reduce incoherent noise present in the observed audio mix. In a general audio coding context, the separation of dominant sources (even mixed with ambience) from background ambiences could be seen as a pre-processing step before applying a dedicated coding scheme. Indeed, the transmission of encoded dominant sources already provides a basic audio scene of the original input.

Considering such multichannel audio signals, next sections are addressing PCA of stereo signals. From the multichannel audio model definition, multichannel audio signal could be naturally considered as multiple stereo signals. Moreover, the multichannel audio model considers time domain signals. Nevertheless, next

sections expose a comparison between time-domain and frequency-domain PCA of stereo signals.

3.2. Time domain PCA of stereo signals

Principal Component Analysis of a stereo signal following the previously described model consists in diagonalizing its channels covariance matrix. The eigenvalue decomposition of the covariance matrix is the first step to be accomplished. From the eigenvalues repartition, PCA consists in projecting stereo channels on the covariance matrix eigenvectors basis. Components resulting from PCA are decorrelated with an energy level proportional to the estimated eigenvalues. Indeed, PCA is known as the optimal decorrelation method which also achieves power concentration.

3.2.1. Repartition of stereo signals eigenvalues

The repartition of stereo signals eigenvalues is an indicative measurement of the input components i.e. directional sources and ambiences, distribution into the output signals resulting from PCA. Stereo signals eigenvalues are estimated from the stereo channels covariance matrix. Using equations (7) and (8), the covariance matrix of stereophonic signals (*M=2*) following the model can be written as:

$$R_{S_2} = \begin{pmatrix} \sum_{d=1}^{D} g_{1,d}^2 \sigma_{S_d}^2 + \sigma_{A_1}^2 & \sum_{d=1}^{D} \left(g_{1,d} \cdot g_{2,d}\right) \sigma_{S_d}^2 + \rho_{A_1 A_2} \sqrt{\sigma_{A_1}^2 \sigma_{A_2}^2} \\ \sum_{d=1}^{D} \left(g_{1,d} \cdot g_{2,d}\right) \sigma_{S_d}^2 + \rho_{A_1 A_2} \sqrt{\sigma_{A_1}^2 \sigma_{A_2}^2} & \sum_{d=1}^{D} g_{2,d}^2 \sigma_{S_d}^2 + \sigma_{A_2}^2 \end{pmatrix} \quad (9)$$

where $\rho_{A_1,A_2} = \dfrac{E\left[A_1 \cdot A_2^T\right]}{\sqrt{\sigma_{A_1}^2 \sigma_{A_2}^2}}$ is the correlation coefficient between ambience signals A_1 and A_2, σ_x^2 is the variance of signal *x*, i.e. $\sigma_x^2 = \mathrm{E}\left[\left(x - \mathrm{E}(x)\right)\left(x - \mathrm{E}(x)\right)\right]$. This covariance matrix is obtained under the assumption that the directional sources are decorrelated: $\rho_{S_i S_j} \approx 0 \ \forall i \neq j$ and $0 < i, j \leq D$.

Eigenvalues can be then computed from the covariance matrix which can be simply written as:

$$R_{S_2} = \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \quad (10)$$

$$\text{then: } \lambda_{1,2} = \frac{1}{2}\left(A + C \pm \sqrt{(A-C)^2 + (2B)^2}\right)$$

$$\lambda_{1,2} = \frac{1}{2}\left(\sum_{d=1}^{D}\left(g_{1,d}^2 + g_{2,d}^2\right)\sigma_{S_d}^2 + \sigma_{A_1}^2 + \sigma_{A_2}^2\right) \pm \frac{1}{2}\sqrt{\left(\sum_{d=1}^{D}\left(g_{1,d}^2 - g_{2,d}^2\right)\sigma_{S_d}^2 + \sigma_{A_1}^2 - \sigma_{A_2}^2\right)^2 + \cdots \left(2\sum_{d=1}^{D}\left(g_{1,d} \cdot g_{2,d}\right)\sigma_{S_d}^2 + 2\rho_{A_1 A_2}\sqrt{\sigma_{A_1}^2 \sigma_{A_2}^2}\right)^2} \quad (11)$$

So, eigenvalues expressions depend on panning gains, variances of directional sources and ambience signals, and the correlation coefficient of ambience signals. λ_1 has been arbitrary chosen as the highest eigenvalue.

The repartition of stereo signals eigenvalues can now be analyzed. Indeed, from real speech and music (instruments) samples, directional sources have been spatialized using conventional amplitude panning technique. Moreover, weakly correlated ambient audio contents, from originals stereo recordings, have been summed with these directional sources. For instance, we have generated a synthetic stereo signal (M=2) constituted of two directional sources (D=2): one speech signal S1 with perceived azimuth varying from left to right speaker and a glockenspiel signal S2 with perceived azimuth varying from right to left speaker (see Figure 3).

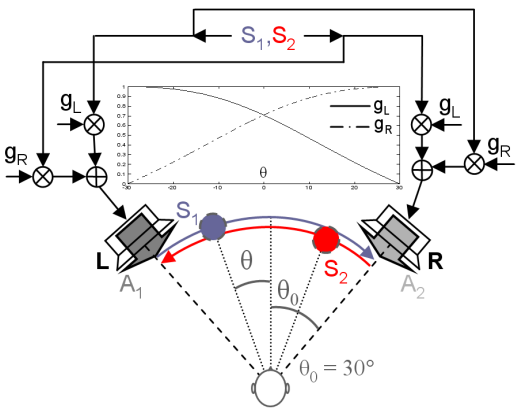

Figure 3: synthetic stereo signal constituted of directional sources (S_1= speech, S_2= glockenspiel), weighted by panning gains (g_L and g_R) and then, summed with weakly correlated stereo ambience signal (A_1, A_2).

Directional sources S_1 (glockenspiel) and S_2 (speech) are weighted by panning gains g_L and g_R following stereophonic law of sines, and then summed with weakly correlated ambience signals A_1 and A_2 coming from a real airport background stereo recording.

$$\begin{cases} L = g_L \times S_1 + g_R \times S_2 + A_1 \\ R = g_R \times S_1 + g_L \times S_2 + A_2 \end{cases} \quad (12)$$

These signals mixing method gives an *a priori* knowledge of panning gains $g_{m,d}$, correlation coefficient of the ambience signals $\rho_{A_1 A_2}$ and powers of directional sources $\sigma^2_{S_d}$ and ambience signals $\sigma^2_{A_1}$ and $\sigma^2_{A_2}$. Original directional sources and ambience signals powers are plotted on Figure 4-(a)-(b). Moreover, we define the directional sources to ambience ratio (DSAR) as the power ratio of the mean power of the directional sources and the mean power of the ambience signals. Actually, we consider ambience signals with equivalent mean power ($\sigma^2_{A_1} \approx \sigma^2_{A_2}$). *DSAR* can be written as:

$$DSAR = 20 \times \log_{10} \left(\frac{\frac{1}{D} \sum_{d=1}^{D} \sigma^2_{S_d}}{\frac{1}{M} \sum_{m=1}^{M} \sigma^2_{A_m}} \right) \text{dB} \quad (13)$$

Considering the original signals used to generate the stereo signal, the *DSAR* estimated from the mean power of directional sources and ambience signals is equal to 30 dB. The mean power of directional sources is equal to -50 dB (see Figure 4-(a)) and the mean power of ambience signals is equal to -80 dB (see Figure 4-(b)). Then, eigenvalues of this stereo signal are estimated according to equation (11) and plotted on Figure 4-(c)-(d).

A comparison of estimated eigenvalues and the *a priori* knowledge of directional sources and ambience signals powers is addressed on Figure 4. The estimated highest eigenvalue λ_1 is equivalent to the power of the dominant directional source: $P_{S_{\max}} = 20 \times \log_{10} \left(\max \left(\sigma^2_{S_1}, \sigma^2_{S_2} \right) \right)$. The small power overhead corresponds to the power of the ambience which spatially coincides with this dominant source. The smallest eigenvalue λ_2 is comparable with the mean power of ambience signals: $P_{A_{mean}} = 20 \times \log_{10} \left(\frac{\sigma^2_{A_1} + \sigma^2_{A_2}}{2} \right)$, when directional sources are spatially coincident (180 < frame number < 280). Moreover, λ_2 is higher than the mean power of ambience signals when directional sources are not spatially coincident (frame number < 180 and frame number > 280). Indeed, λ_2 is comparable with the power of secondary directional source plus the power of ambience which spatially coincides with this secondary source.

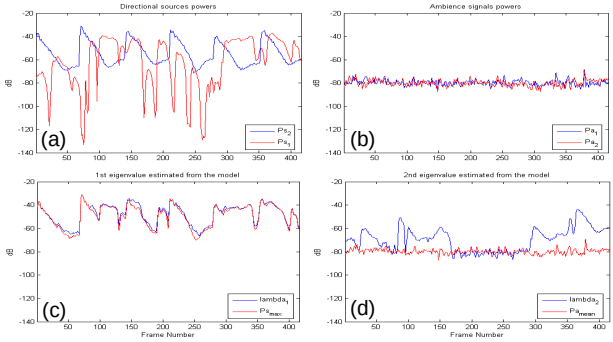

Figure 4 - (a): original directional sources (S_1=speech and S_2=glockenspiel) powers. (b): original ambience signals (stereo airport recording) powers. (c): 1st eigenvalue estimated from the model is slightly higher than the power of the dominant directional source. (d): 2nd eigenvalue estimated from the model is generally higher than the mean power of ambience signals. This power overhead corresponds to the power of the secondary directional source.

We synthesized several stereo signals with the same signals (sources and ambience) but with different *DSAR* varying from 5 to 50 dB. Indeed, the ambience signals are weighted by a coefficient which results in the desired *DSAR*. Actually, *DSAR* decreases, with 5 dB step, when the mean power of ambience signals increases (see Figure 5). Estimated eigenvalues for all stereo signals with different *DSAR* are plotted on Figure 5 which addresses a comparison between these estimated eigenvalues and the power of the dominant directional source ($P_{S_{\max}}$) - see Figure 5-(a)) - and the mean power of the ambience signals ($P_{A_{mean}}$) - see Figure 5-(b).

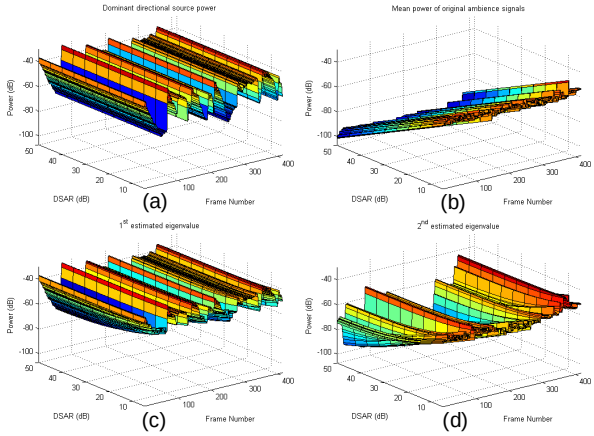

Figure 5: the four draws corresponds to signals powers which are function of time (frame number) and DSAR (from 5 to 50 dB). (a): power of the original dominant directional source – (b): mean power of the originals ambience signals – (c): power of the 1st estimated eigenvalue – (d): power of the 2nd estimated eigenvalue.

The estimated highest eigenvalue λ_1 is equivalent to the power of the dominant directional source plus the power of the ambience which spatially coincides with this dominant source. Moreover, Figure 4-(c) shows that this ambience power increases when the *DSAR* decreases. The smallest eigenvalue λ_2 is comparable with the mean power of ambience signals plus the power of secondary sources. Naturally, lower is the *DSAR* and more the ambience power recovers these secondary sources powers.

Starting from stereo signals eigenvalues repartition, next section addresses the PCA transformation based on eigenvalues repartition also known as Karhunen-Loève Transform (KLT).

3.2.2. Time domain PCA by rotating stereophonic signals

Time domain PCA consists in projecting the data on the basis of stereo signal covariance eigenvectors. The computation of eigenvectors matrix *V*, of dimension 2x2 (in case of 2 channels signal), is performed according to the eigenvalues estimation and the covariance matrix R_{S_2} diagonalization:

$$V \cdot R_{S_2} \cdot V^T = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \text{ where } R_{S_2} = \begin{pmatrix} A & B \\ B & C \end{pmatrix} \qquad (14)$$

where: $\lambda_i, i \in [1;2]$ are the stereo signal eigenvalues such as $\lambda_1 > \lambda_2$. Orthogonal eigenvectors allow the use of a rotation matrix which can be written as:

$$V = \Re(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \qquad (15)$$

where the rotation angle expression is derived from the diagonalization of the covariance matrix (equation 14):

$$\theta_1 = \arctan\left(\frac{\lambda_1 - A}{B}\right), \quad \theta \in \left[-\frac{\pi}{2};\frac{\pi}{2}\right] \qquad (16)$$

Moreover, λ_1 can be written as a function of the covariance matrix elements (A, B and C), see equation (10). Then, from equations (16) and (10), another expression of θ which only needs the estimation of the covariance matrix is:

$$\theta_2 = \frac{1}{2}\arctan\left(\frac{2\times B}{A-C}\right), \quad \theta \in \left[-\frac{\pi}{4};\frac{\pi}{4}\right] \qquad (17)$$

Therefore, PCA transformation also known as KLT is achieved by rotating the stereophonic data. From the original correlated stereo signal, this rotation results in one principal component *PC* and one remaining signal *A* for each block *n* (windowed stereo data).

$$\begin{pmatrix} L(n) \\ R(n) \end{pmatrix} \cdot \begin{pmatrix} \cos\theta_n & \sin\theta_n \\ -\sin\theta_n & \cos\theta_n \end{pmatrix} = \begin{pmatrix} PC(n) \\ A(n) \end{pmatrix} \qquad (18)$$

PCA drastically reduce the correlation of the original stereo signal i.e. PC and A components are de-correlated - see the covariance matrix of the rotated signals on equation (19):

$$\begin{aligned} R_{PC,A} &= \mathrm{E}\left[\left(\begin{pmatrix} L \\ R \end{pmatrix} \cdot \Re(\theta)\right) \cdot \left(\begin{pmatrix} L \\ R \end{pmatrix} \cdot \Re(\theta)\right)^T\right] \\ &= \Re(\theta) \cdot \mathrm{E}\left[\begin{pmatrix} L \\ R \end{pmatrix} \cdot \begin{pmatrix} L \\ R \end{pmatrix}^T\right] \cdot \Re^T(\theta) \\ &= \Re(\theta) \cdot R_{S_2} \cdot \Re^T(\theta) \\ &= \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \end{aligned} \qquad (19)$$

Moreover, PCA achieves an energy concentration to the principal component according to the value of λ_1. From the eigenvalues repartition analysis, the remaining signal could be denoted as ambience signal with low energy according to the value of λ_2.

The overlap-and-add (OLA) method is used to generate the final signals *PC* and *A*. As a result, the estimation of θ is the main operation to realize the PCA of any stereo signal. Besides, this estimation should provide a constant sign of the rotation matrix. Indeed, if rotation matrices of adjacent blocks have opposite sign then the OLA synthesis would results in losses of information. Then, to achieve a robust analysis /synthesis scheme, the rotation angle should belong to the interval $[0;\pi/2]$ and then provide a constant positive sign of the rotation matrices. The expressions (16) and (17) provide estimations of the rotation angle which could be negative. These negatives values of θ should be avoided without modifying its the real physical sense i.e. the dominant source azimuth in the stereo image $[-\pi/6;\pi/6]$. Actually if we consider an estimation of the rotation angle belonging to $[0;\pi/2]$, this interval may be put in correspondence with the real stereo image interval. Then, a dominant source with a real azimuth equal to -30° (respectively +30°) should results in rotation angle estimation equal to 0° and then provide *PC(n) = L(n)* (respectively +90° and then provide *PC(n) = R(n)*).

Several modifications of the estimated rotation angle are possible to avoid losses of information during the OLA synthesis. The first one, may be the most natural, could be written as:

$$\tilde{\theta}_i = \begin{cases} \theta_i, & \text{if } \theta_i \geq 0 \\ \theta_i + \dfrac{\pi}{2}, & \text{else} \end{cases}, \quad i \in [1;2] \qquad (20)$$

Therefore, $\tilde{\theta}_i \in [0;\ \pi/2]$ and $\tilde{\theta}_1 = \tilde{\theta}_2$. But this operation does not conserve the real azimuth of the dominant source. Indeed, a dominant source with a real azimuth equal to 29° (on the right of the stereo image [-30°;+30°]) should be equivalent to an ideal estimate of the rotation angle equal to 88.5°. Actually, such a stereo signal with weakly correlated ambiences summed with this one-sided (right side) dominant source could have close to zero and also negative cross-correlation (*B*). Then, equation (16) could provide an estimated rotation angle θ_1 equal to +/-88.5° depending on the sign of the channels cross-correlation. Indeed, $(\lambda_1 - A)$ in equation (16) is always positive. So, close to zero and negative cross-correlation provides estimated value of θ_1 equal to -88.5°. The corrected estimation is then $\tilde{\theta}_1 = +1.5°$ which does not yield a correct estimation of the rotation angle. If we consider the equation (17) with the same example i.e. with a close to zero and negative cross-correlation (*B*) and the negative difference (*A*-*C*) – the dominant source is one-sided on the right channel then: $A<C$ – the estimated θ_2 is equal to +1.5° and then will not be corrected.

This analysis shows how negative cross-correlation can alter the estimated values of θ. Another approach consists in limiting the cross-correlation values to a minimum equal to zero i.e. negatives values of cross-correlation are set to zero. Unfortunately, this leads to only use the equation (17) which would provide some false estimated values of θ_2 equal to 0° although the dominant source could be located at many azimuths corresponding to the difference *A*-*C*.

As a result, the only solution which satisfies our goal is to consider the absolute value of the cross-correlation. Then, the rotation angle estimated from equation (16) i.e. θ_1, could be directly put in correspondence with real azimuth of the dominant source. The rotation angle estimated from equation (17) i.e. θ_2, needs to correct the estimated negatives values (when $A<C$) even with absolute values of cross-correlation. Then the estimated rotation angle expressions of $\hat{\theta}_1 = \hat{\theta}_2$ can be summarized as:

$$\begin{cases} \hat{\theta}_1 = \arctan\left(\dfrac{\lambda_1 - A}{|B|}\right), & \theta_1 \in \left[0;\dfrac{\pi}{2}\right] \\ \hat{\theta}_2 = \begin{cases} \dfrac{1}{2}\arctan\left(\dfrac{2\times|B|}{A-C}\right), & \text{if } A-C \geq 0 \\ \dfrac{1}{2}\arctan\left(\dfrac{2\times|B|}{A-C}\right)+\dfrac{\pi}{2}, & \text{else} \end{cases}, & \theta_2 \in \left[0;\dfrac{\pi}{2}\right] \end{cases} \qquad (21)$$

From the stereo signal exposed at section 3.2.1 (see Figures 3 and 4), the estimated rotation angles referring to equations (20) and (21) are plotted on Figure 6.

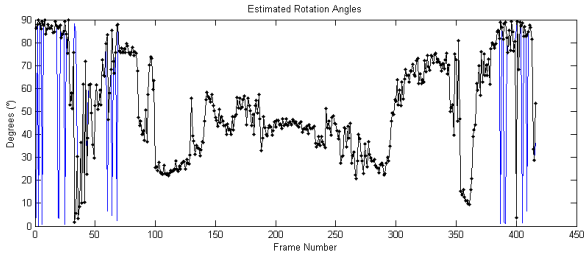

Figure 6: estimated rotation angles from sine windowed and 50% overlap blocks analysis. The plotted solid line corresponds to the estimate $\tilde{\theta} = \tilde{\theta}_1 = \tilde{\theta}_2$ and the plotted dotted line corresponds to the estimate $\hat{\theta} = \hat{\theta}_1 = \hat{\theta}_2$.

The rotation angle estimated from the original covariance computed for each block has values corresponding to the azimuth of the dominant directional source for each block. Indeed, the first values of $\hat{\theta}$ are close to 90° which means that the directional source located on the right of the stereo image (i.e. the glockenspiel) is considered as the dominant source. This observation is confirmed by the originals directional sources powers on Figure 4-(a). The next values of the estimated $\hat{\theta}$ are close to 0° (30<Frame Number<50) which means that the dominant source is located on the left of the stereo image as it is confirmed by the Figure 4-(a) i.e. the speech source at the left of the stereo image as a higher power than the glockenspiel source for these frames. Moreover, when the directional sources spatially coincides (200<Frame Number<220) at the middle of the stereo image, the estimated rotation angle $\hat{\theta}$ is naturally close to 45°. The first values of the estimated rotation angle $\tilde{\theta}$ are discontinued and rock between the minimum and the maximum values of the interval [0;90]° as the previous analysis has predicted it. Besides, the same observation can be addressed for the last analyzed blocks (Frame Number > 370). Indeed, from these analyzed blocks (70 > Frame Number > 370), negative cross-correlation occurs when the directional sources are located at extremes azimuths.

Finally, a robust and effective estimation of the rotation angle needed to achieve PCA/KLT is expressed by the equation (21) which permits an estimation only based on the covariance matrix elements ($\hat{\theta}_2$ expression). Then, dominant directional source moving in the stereo image [-30,30]° can be located by an estimated rotation angle varying between [0,90]°.

Actually, PCA of stereo signals following the time-domain model previously presented could be applied to frequency-domain signals. Next section presents subbands eigenvalues repartition and corresponding PCA of stereo signals transformed in the frequency domain. Finally, a comparison of time-domain and frequency-domain PCA is given.

3.3. Frequency subbands vs. time domain PCA of stereo signals

The eigenvalues of stereo signals may also be estimated in the frequency domain. Moreover, if we consider band limited signals in the frequency domain, the eigenvalues estimation will provide several estimates i.e. two eigenvalues per subband, and then provide a thinner analysis.

3.3.1. Frequency subbands separation

The frequency transformation applied to stereo channels is the short time Fourier transform (STFT). The parameters of the STFT used are a sine window of length equal to N=1024 samples, the transform size is also equal to K=1024 frequency bins (no zero-padding) and the frames overlap is 50%. Then, a N_b=20 subbands rectangular frequency windowing, following the ERB scale, is applied to the complex spectra $F_{C(n)}(k) = \left|F_{C(n)}(k)\right| e^{j\phi_{C(n)}(k)}$, with channel block $C(n)$ and k is the frequency index such as: $k \leq f_s/2$, where f_s is the sampling frequency. Then, this process results in N_b frequency subband (with k_b frequency bin start index) spectra $F_{C(n,b)}(k_b, ..., k_{b+1} - 1)$ for each frame n.

Next section addresses the benefit of frequency subbands eigenvalues estimation vs. time domain eigenvalues estimation. More precisely, we argue this benefit only if no spatially coincident directional sources have different frequency supports.

3.3.2. Frequency subbands eigenvalues repartition

From the same stereo signal used in section 3.2.1, a subbands analysis of its directional sources and ambience signals has been achieved (see Figure 7).

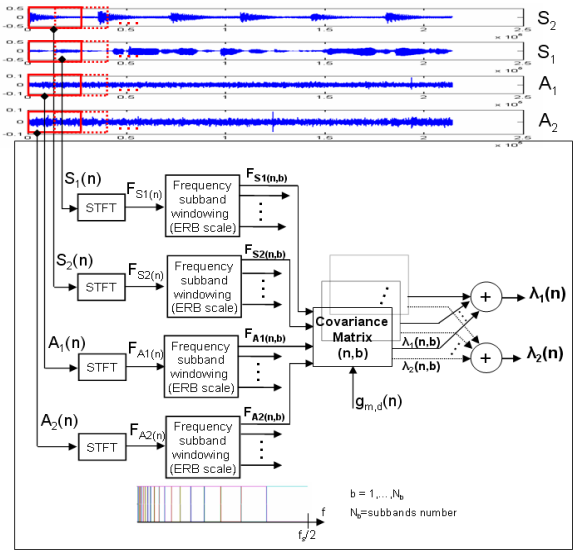

Figure 7: stereo signal eigenvalues estimated from directional sources and ambience signals frequency subbands spectra. The covariance matrix is estimated for each frame n and each subband b.

The covariance matrix of the subband spectra is estimated according to equation (9). More precisely, there is an equivalence of centred signals powers in time and frequency domains which can be written as:

$$\sigma^2_{X^c}(n,b) = \frac{2}{NK} \cdot \sum_{k=k_b}^{k=k_{b+1}-1} \left| F_{X^c(n)}(k) \right|^2 \qquad (22)$$

where $F_{X^c(n)}$ is a the STFT of block signal $X^c(n)$ i.e. a vector of size $K/2+1$ considering the spectral hermitian symmetry. The cross-correlation of centered ambience signals is estimated from the centred ambience subband cross-spectrum, with frequency index k, as:

$$\rho_{A_1^c A_2^c}(n,b) = \frac{1}{N} \times \Re\left(\frac{2}{K} \cdot \sum_{k=k_b}^{k=k_{b+1}-1} F_{A_1^c(n)}(k) \cdot F^*_{A_2^c(n)}(k) \right) \qquad (23)$$

Then, the estimated covariance matrix from the original centered signals subband spectra ($F_{S_1^c(n)}$, $F_{S_2^c(n)}$, $F_{A_1^c(n)}$, $F_{A_2^c(n)}$) is used to estimate the subband eigenvalues according to equation (11) – see Figure 7.

The eigenvalues estimated from the temporal analysis are then compared with the sum of eigenvalues estimated from the subbands analysis (see Figure 7): $\sum_{b=1}^{N_b} \lambda_i(n,b), \quad i \in [1;2]$.

The comparison between the eigenvalues estimated in time and frequency domains is addressed at Figure 8. The lowest eigenvalue estimated from subbands frequency analysis does also corresponds to secondary directional sources plus mean ambience signals power as the lowest eigenvalue estimated from time domain analysis (see Figure 8-(b)). Moreover, some directional sources considered as secondary sources with the time domain analysis have been considered as dominant directional sources with the frequency subbands analysis. Indeed, the subband analysis provides a thinner analysis which results in one dominant source per subband. Then, when dominant and secondary (considering time domain approach) sources have different frequency support, the secondary source can be considered as dominant in the considered subband. Although the analysis complexity has increased, compared to time domain analysis, the lowest eigenvalue has a power much closer to the original ambience signals mean power.

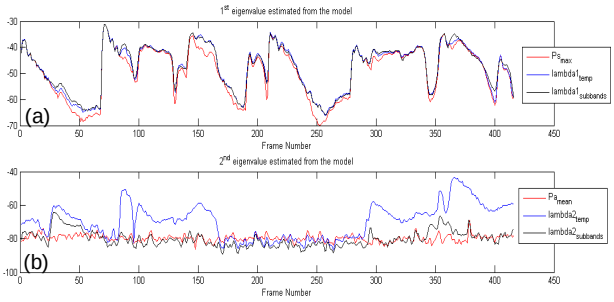

Figure 8: (a): 1st eigenvalue estimated from temporal or subbands frequency analysis are equivalents. (b): 2nd eigenvalue estimated from subbands frequency analysis is closer to the mean power of ambience signals than the eigenvalue estimated from time domain analysis.

The estimated eigenvalues from subbands covariance matrix of stereo signals following the model could be also estimated directly from any stereo signal. Thus, the subbands covariance matrix of the stereo signal is estimated without any knowledge of the directional sources and ambience signals.

3.3.3. Subbands PCA of stereophonic signals

As we expose subbands eigenvalues estimation in section 3.3.2, we can derive PCA of stereo signals transformed in the frequency domain. Therefore, frequency PCA processing is based on the covariance matrix computed for each frequency subband:

$$R_{S_2}(n,b)=\begin{pmatrix} A(n,b) & B(n,b) \\ B(n,b) & C(n,b) \end{pmatrix}, \text{ with:}$$

$$\begin{cases} A(n,b)=\dfrac{2}{NK}\cdot\displaystyle\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{X_L^c(n)}(k)\right|^2 \\ B(n,b)=\dfrac{2}{NK}\times\Re\left(\displaystyle\sum_{k=k_b}^{k=k_{b+1}-1}F_{X_L^c(n)}(k)\cdot F^*_{X_R^c(n)}(k)\right) \\ C(n,b)=\dfrac{2}{NK}\cdot\displaystyle\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{X_R^c(n)}(k)\right|^2 \end{cases} \quad (24)$$

The channels auto-covariance (*A* and *C*) correspond to the mean power spectral density of the subband spectra. The channels cross-covariance (*B*) is estimated from the cross-spectrum of the stereo channels.

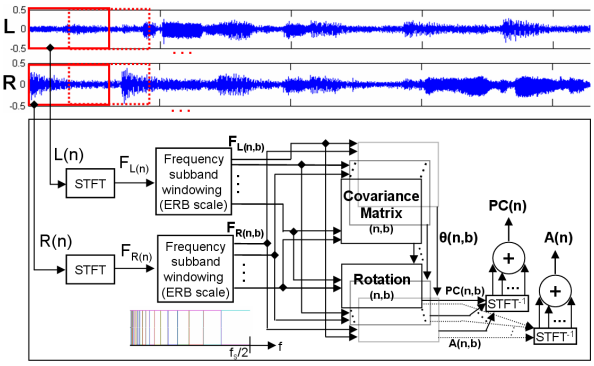

Figure 9 : frequency subbands PCA applied to a stereophonic signal (L, R). PCA is achieved by rotating the stereo subbands and results in one principal component PC and one ambience signal A.

As for time domain PCA, the rotation angle $\hat{\theta}$ is computed for each subband according to equation (21) where all quantities are estimated in the frequency domain (per subband). From the stereo signal exposed at section 3.2.1 (see Figure 3 and 4), the estimated rotation angles (in degrees) are plotted on Figure 10 which axes are time (x-axis) and frequency (y-axis).

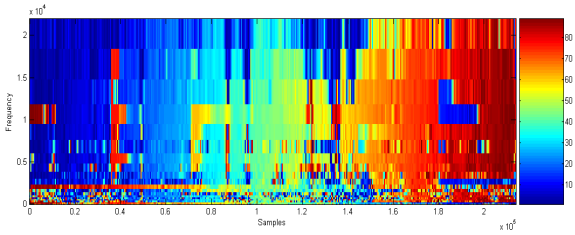

Figure 10: rotation angles (in degrees) estimated from the stereo signals defined at section 3.2.1. The glockenspiel is considered as the dominant directional source only for some subbands comparing to the time domain estimation of the rotation angle (see Figure 6).

So, one dominant directional source is located by the estimated rotation angles for each subband. Then, for each analyzed signal block, as much dominant sources than subbands number are located by the estimated rotation angles. These subbands rotation angles are then used to rotate the subbands original input stereo data. Subband PCA processing results in two frequency components for each subband: the principal component *PC(n,b)* and the ambience component *A(n,b)*. Time domain band limited signals are then obtained with inverse short time Fourier transform ($STFT^{-1}$) applied to these frequency subband components (see Figure 9). Finally, time domain full band signals *PC* and *A* result from the sum of all corresponding band limited signals. This sum is achieved via overlap-add method.

Therefore, a frequency subbands PCA processing results in the extraction of one dominant direct source per subband against only one dominant direct source with a temporal processing. We can now conduct the comparison between time and frequency domains PCA according to the energy of the transformed signals *PC* and *A*. A relevant measurement of the energy compaction into the principal component is achieved by computing the Principal Component to Ambience energy Ratio (PCAR). From eleven miscellaneous stereo signals with sampling frequency equal to 44100 or 48000 Hz, time-domain and subbands frequency PCA has been achieved. Then, PCAR has been computed from the rotated signals of the eleven input signals. The comparison is addressed on Figure 11. Subbands frequency PCA results in a better energy concentration than time-domain PCA does. From eleven stereo signals rotated in time and frequency domains, the mean PCAR difference is about 2 dB.

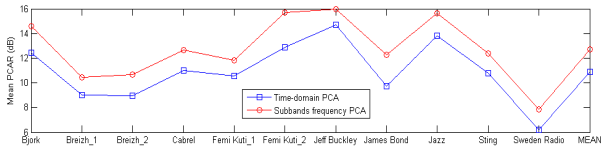

Figure 11: mean PCAR over analyzed frames of 11 principal components and ambiences resulting from time-domain PCA and subbands frequency PCA of the original stereos.

We show how a frequency PCA processing, based on the covariance matrix computed for each frequency subband, could improve the results obtained in the time domain in terms energy compaction into the rotated signals. PCA features can then naturally be used in a data compression context and also for upmix stereo signals to multichannel audio (as described in [6]-[7]). Considering correlated audio inputs, the output signals of frequency subbands PCA constitute a compact representation of the original input. The principal component should be coded and transmitted in order to recover the main information of the original input. Then, to achieve the inverse transformation, the ambience component should be transmitted with a bitrate as high as the audio quality level is desired.

4. PARAMETRIC STEREO CODING BASED ON PCA

Starting from the analysis made in section 3, it is possible to encode a stereo signal with frequency subbands PCA pre-processing. This method has already been proposed in [11] where PCA is obtained by rotating the subbands of the stereo signal. PCA is used as a power concentration processing such as Mid/Side coding scheme which encode the sum and difference signals of stereo channels. Even if PCA features, such as power concentration and full decorrelation, are required for optimal bite rate reduction, traditional encoding of the transformed signals does not yield significant coding gain (see [11]).

In order to provide a low bit rate audio coding method compatible with most networks, a parametric coding of the ambience signal(s) is achieved.

The coding scheme consists in a traditional monophonic coding of the principal component *PC* and a parametric coding method of the ambience signal *A* (cf. Figure 12).

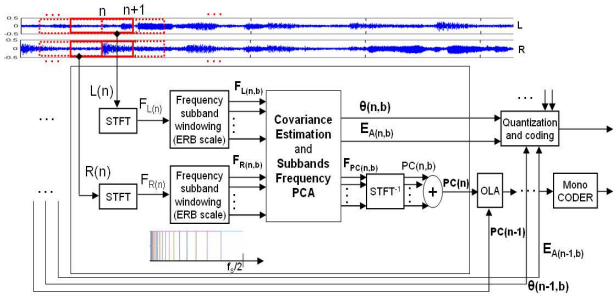

Figure 12: parametric coding of stereo signals based on subbands frequency PCA processing.

Subbands PCA processing is achieved as described in section 3.3. Moreover, the frequency subbands ambience signal resulting from subbands PCA processing is analyzed. This frequency analysis allows the extraction of energy parameters referring to the ambience subbands energies $E_{A(n,b)}$. These energy levels are extracted for each frame *n* and each subband *b* and then quantized and transmitted. Moreover, considering the fact that the ambience signal has weak energy level, the difference between the mean energy of the ambience signal block and the energy subbands values should be even weaker and then quantized.

So, the proposed coding method can be summarized as a monophonic coding of the principal component and the quantification and transmission of the following parameters:

- PCA rotation angles $\theta(n,b)$
- desired subbands energy $E_{A(n,b)}$ (or the difference with the mean energy level) of the ambience signal resulting from PCA

The decoding scheme is based on the generation of an ambience signal *A'*, from the decoded signal *PC'* and the dequantized parameters. Therefore, the inverse PCA can then synthesize a stereo signal perceptually as close as possible from the original stereo.

Due to the PCA property of decorrelation, the decoder should generate an ambience signal weakly correlated to the decoded principal component. However, the frequency synthesis of subbands signal *A'* from the principal component and the energy parameters $E^{Q}_{A(n,b)}$ only provide to *A'* its spectral envelope. To achieve weakly correlation between *PC'* and *A'*, we propose the use of random phase all-pass filters as described in [12]. More precisely, we realize a frequency filtering (filter H

on Figure 13) of subbands signal *A'*. So, the decoder can then realize the inverse subbands PCA from the signals *PC'* and A'_H and the dequantized rotation angles $\theta^Q(n,b)$ (see Figure 13). Afterwards, inverse STFT of the subbands signals obtained from inverse subbands PCA are summed to generate the stereo signal (*L'*, *R'*).

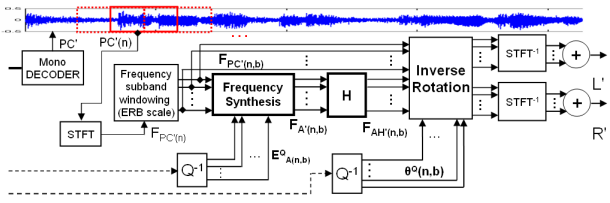

Figure 13: parametric decoding of stereo signals based on inverse subbands PCA processing.

The main advantage of this parametric stereo coding method is the fact that the decoding process is strictly the inverse transformation of the encoding process.

5. CONCLUSION

We have argued that the UDT could be seen as an equivalent of BCC scheme in a stereo coding context. Moreover, a deeper analysis could results in same conclusions considering coding of multichannel audio signals. We have then introduced a multichannel audio model defined as directional sources and ambiences. The eigenvalues analysis of stereo signals has showed that PCA of stereo signals deliver two decorrelated components with energy level corresponding to the eigenvalues repartition. Moreover, frequency subbands PCA yields more efficient power concentration than classical time domain PCA. This frequency subbands analysis scheme leads the possibility to realize audio coding at low data rate. Indeed, a parametric coding method of stereo signals based on PCA is exposed. Next investigations are directed towards the quantization of parameters which will deliver resulting bitrates of the stereo codec. Then, listening tests will determine the interest of such a coding method. Finally, PCA based extensions for parametric coding of multichannel audio signals need to be investigate.

6. REFERENCES

[1] R. Dressler, "Dolby surround Pro Logic decoder principles of operation".

[2] C. Faller, "Parametric Coding of Spatial Audio", PhD Thesis, EPFL, 2004.

[3] B.R. Glasberg, B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data", Hearing Research 47, 103-138.

[4] J. Breebaart and al., "Parametric Coding of Stereo Audio", EURASIP Journal on Applied Signal Processing 2005:9, 1305-1322.

[5] K.M. Short & al, "Multi-Channel Audio Processing Using a Unified Domain Representation", Proc. 119th AES convention, New York, October 2005, Preprint 6526.

[6] R. Irwan & R.M. Aarts, "Two-to-Five Channel Sound Processing", JAES, Vol. 50, No. 11, 2002.

[7] Y. Li & P.F. Driessen, "An Unsupervised Adaptive Filtering Approach of 2-To-5 Channel Upmix", Proc. 119th AES convention, New York, October 2005, Preprint 6611.

[8] Producers & Engineers Wing Surround Sound Recommendations Committee, "Recommendations for Surround Sound Production", the National Academy of Recording Arts & Sciences, 2004.

[9] N. Mitianoudis and M. Davies, "Audio Source Separation: Solutions and problems", Int. J. Adapt. Control Signal Process, 18:299-314, 2004.

[10] F. Asano and al., "Effect of PCA filter in blind source separation", Proc. ICA2000, pp.57-62.

[11] R.G. van der Waal & R.N.J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals", International Conference on Acoustics, Speech, and Signal Processing, Toronto, 1991.

[12] M. Bouéri & C. Kyirakakis, "Audio Signal Decorrelation Based on a Critical Band Approach", Proc. 117th AES convention, San Francisco, October 2004, Preprint 6291.