CHALMERS
UNIVERSITY OF TECHNOLOGY

Stereo-to-Five Channels Upmix Methods

Implementation and Comparative Study

Master's thesis in MSc Sound and Vibration

PANTELEIMON PAPASTERGIOU

Department of Architecture and Civil Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Master's thesis BOMX60-18-1

Stereo-to-Five Channels Upmix Methods

Implementation, Objective and Subjective Evaluation of Primary-Ambience Extraction

PANTELEIMON PAPASTERGIOU

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2018

Stereo-to-Five Channels Upmix Methods

Implementation and Comparative Study
Panteleimon Papastergiou

Supervisor: Jens Ahrens, Chalmers University of Technology and Jonatan Ewald, Volvo Cars
Examiner: Jens Ahrens, Department of Architecture and Civil Engineering

Master's Thesis BOMX60-18-1
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Brian Eno, The Ship, 2014-2016. Installation view from Kunsthal Charlottenborg, 2016.

Typeset in LaTeX
Printed by [—]
Gothenburg, Sweden 2018

Stereo-to-Five Channels Upmix Methods
Implementation and Comparative Study
Panteleimon Papastergiou
Department of Architecture and Civil Engineering
Chalmers University of Technology

Abstract

The aim of several 3D audio concepts and products is to create a more immersive, engaging and natural-sounded listening experience. Emerging audio signal processing techniques, make it possible for regular stereo recordings to be compatible and reproduced with multichannel home theatre or automotive loudspeaker audio systems.
In this thesis, various existing methods are investigated and implemented for converting stereo recordings to four or five channels in the primary-ambience extraction (PAE) framework. In that, audio signals are often considered linear combinations of primary and ambient components. The former are assumed to be correlated, whereas the latter uncorrelated. The basic function of the upmix systems is to remove the correlated components from the electronic audio material, which are intended for playback with loudspeakers behind the listeners, in a 3/2 or 2/2 configuration. That way the decomposition facilitates the appropriate rendering for spatial enhancement.
The upmixers, either keep the initial stereo recording in the frontal loudspeakers or add a third central channel in the frontal setup to allow for off the "sweet spot" listening. All the methods are implemented in the frequency domain using the widely known short time Fourier transform (STFT) technique, except one. Central in the development of the algorithms in frequency domain are the method of Principal Components Analysis (PCA), the least squares estimates (LS), the normalized least mean squares (NLMS) adaptive filter and certain ambience masking functions. On the other hand, the core of the only time domain method is the least mean squares (LMS) adaptive filter.
Assessment of the new upmix systems was accomplished in an objective and subjective way; firstly, using performance measures such as the ambience energy fraction (E_A) and the cross-correlation coefficient of primary and ambient components (ϕ_P and ϕ_A respectively), and secondly with a listening test which requires from the participants to judge the systems according to the overall impression.
The objective and subjective evaluation results suggest that a subjectively tuned ambience masking function and the frequency domain NLMS algorithm provide both promising upmix solutions and computational advantage.

Keywords: spatial audio, spatial enhancenment, upmix, primary ambience extraction, adaptive filtering, .

Acknowledgements

First of all I would like to thank my supervisors Jens Ahrens at Chalmers and Jonatan Ewald at Volvo, for the helpful advice and productive collaboration and of course for the fruitful discussions. Deep thanks to my family and all my good friends for their endless support and love.

Panteleimon Papastergiou, Gothenburg, January 2018

Contents

List of Figures

List of Tables

1
Introduction

The concept of reproducing sound with more than two loudspeaker setups is not novel. It has been a while since the first time surround reproduction systems have been installed for domestic and automotive use, with the 5.1 setup being the most frequently used [1]. The evolving technology has also yielded audio signal transmission with broad spectral bandwidth, highly effective spatial sound recording techniques and configurations incorporating multiple loudspeakers. For instance, in 7.1 the angle is decreased among loudspeakers, whereas as mentioned in [2], elevated loudspeakers are introduced in systems like 9.1 11.1. In addition, the fast paced development of virtual reality video has emerged the need for 3-D audio reproduction. Spatial sound has also gained ground in critical applications associated with route guidance of people having seriously defective vision [3] and with flight navigation systems [4].

It is apparent that, spatial sound systems have to be able to interact with different audio channel formats, because it's not always technically-and financially-advantageous to use recording techniques, typical for every reproduction system [11]. Thus, compatibility turns out to be a very important issue. Besides, the vast majority of commercial music is recorded in stereo format. Therefore, certain signal processing techniques have to be employed, in order to convert the stereo recordings to multichannel formats, maintaining the quality of audio information and enhancing the listening experience.

Extensive research has been conducted for spatial audio processing [6], [7] and spatial audio coding, which comprises spatial audio scene coding (SASC) [5] and directional audio coding (DirAC) [8]. The aim is to decompose the input audio signal into non-diffuse and diffuse sound. The calculated diffuse sound in the channels differs from positive and negative correlations, which deviates from the ambience definition given in the primary-ambience extraction framework investigates in the present thesis. In SASC, the the Gerzon localization vector [9] is used to carried out localization analysis, separately for both the separated components. In DirAC, the primary sound is rendered applying vector base amplitude panning (VBAP) [10], while the diffuse sound is decorrelated and played back by rear loudspeakers. The aim for these signal processing techniques is to achieve spatial sound reproduction with every sound system configuration.

In [12] He refers to a broad category of methods, which attempts to solve the compatibility mismatch of audio channels with different loudspeaker numbers, as audio-remixing. It includes two subcategories producing the inverse results; the downmix and upmix of audio signals. The former class of methods reduces the number of channels, whereas the latter creates surround information and in some cases also

extracts a central front channel for off the "sweet" spot listening, recalculating the weights and the energy [13].
The downmix process is employed to create 3-D audio environments radiated from stereo setups, as discussed in [14]. Such applications are very useful in portable audio devices, like cellphones and mp3 players. Moreover, in [15] a downmixed audio channel, along with spatial cues, is used for multichannel audio rendering.
On the other hand, upmix is the enhancement of sound reproduction, creating listening environments, which essentially incorporate spatial properties extracted from and added to the stereo format. In a framework of upmix, a sound event is considered as a combination of primary (or direct) and ambient (or reverberant) component [17]. The primary signal is an aggregation of directional sources, while the ambience is the diffuse part, ideally coming from all directions with the same energy [44].
At first, it was Orban who developed a method for extracting pseudo-stereo from monophonic signal [16]. To date, the upmix concept has been further evolved by several methods, based on primary-ambience separation. For instance, in [17] a time-frequency mapping function identifies and masks the ambient component, assuming that the ambience has equal ratios to the initial stereo channels. A similar method in [19], differently defines the mask function, assuming equal levels of ambience. However, principal component analysis (PCA) remains the most broadly investigated with [21], [22] and [20] as examples. According to [23], the idea is to reduce the dimensions of an input's data set, by creating a new one with orthogonal components, retaining most of the variation of the initial data. It is further assumed, that the principal components have significantly higher amplitude than the ambience in the channels. Another method for primary-ambience separation is described in [13]. It is based on calculation of the least squares (LS) estimates of the signal components, while the normalized least mean (NLMS) squares are employed in [24].
That's exactly the first stage of the master thesis work; to study the existent literature for primary-ambience decomposition and to identify the most promising stereo-to-five (or four) upmix approaches. The next steps are the implementation of the identified methods and the perceptual evaluation (user study) according to listeners' preference, of each one of them . The latter also constitutes **the fundaental aim of the project, which is the elicitation of an inference in respect to the listeners' greater preference for an upmix method, reproduced by the fixed 5.1 loudspeaker setup**. Of course, for methods which have alternative versions, like PCA, a representative one resulting in the optimal separation is implemented here.
In a nutshell, the thesis is structured as follows: the second chapter introduces the most important aspects related to human hearing, while chapter three contains the mathematical background, of the primary-ambience extraction framework. In chapter four the upmix methods are described and in five the ambience extraction algorithms are assessed by simulations with objective measures, while in chapter six the subjective evaluation results are presented and discussed. In the end, chapter seven concludes this thesis work and points out possible future work.

2

Fundamentals of Spatial Hearing

This chapter presents the thesis background related to binaural hearing. Since, the upmix conversion is reproduced by a 5.1 surround system, the estimated dimension of the sources is the lateral one. Thus, the fundamentals of localization cues of human hearing mechanism and the localization of sources on the horizontal plane are briefly described in the forthcoming sections.

2.1 Binaural Localisation

2.1.1 HRTF

The sound events take place in the three dimensional space. The most common coordinate system are the cartesian and the polar coordinate system. They determine the definition of any physical event. Sound events are better described by the polar or spherical system. The dimensions of distance, azimuth and elevation are used for the particular context. Figure 2.1 depicts the dimensions, considering the centre human head as the centre of this imaginable sphere. Distance (r) is the direct path from the center of the head to the sound event. Azimuth (ϕ) is the angle between the source at 0° and at a random position on the horizontal plane, whereas elevation (θ) is the equivalent on the median plane.

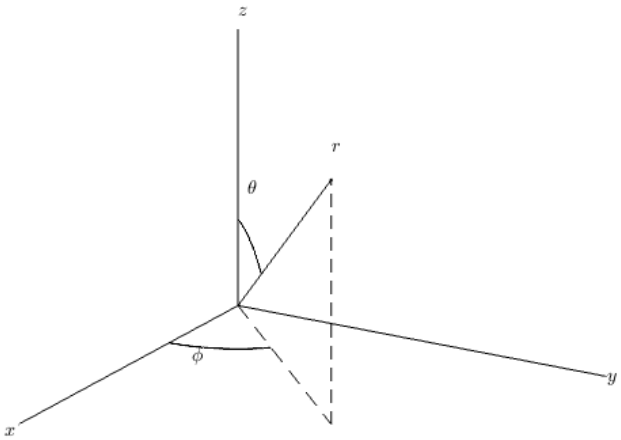


Figure 2.1: Spherical coordinates. The start of the axes is considered to be the center of the head.

The human ability to localize sounds is based on the different paths the sound follows to reach every ear. The head-related transfer functions (HRTF) describe the spectrum filtering, caused by the interactions of the waves with the head, torso and pinna. More specifically, the HRTF is defined as the ratio of the sound pressure at the ears, to the pressure measured by a microphone placed at the position representing the middle of the head, in the free field [25]. Since, these functions depend on the frequency and the polar coordinates, for both ears in frequency domain it applies that,

$$H_L(f) = \frac{P_L(f)}{P_0(f)}, \tag{2.1}$$

$$H_R(f) = \frac{P_R(f)}{P_0(f)} \tag{2.2}$$

where f denotes the frequency.

As it may be implied, the HRTFs also depend on the different properties of the head and torso of the people. In time domain, one refers to the head related impulse response (HRIR) [12].

2.1.2 Interaural Cues

Regarding the types of localization cues, there are monaural and interaural ones. The monaural cues are interpreted by the human brain as information only at one ear and they are associated with the identification of the distance of the source, the elevation angle and the median plane. On the other hand, the interaural cues are related with the signal differences at both ears. They are linked to the azimuth localisation.

Interaural time difference (ITD) and the interaural level difference (ILD) are introduced by Rayleigh in his duplex theory in [26]. The ITD indicates the difference in time the sound travels to the left and right ears. The sound waves coming from a sound source, which is located sideways to the head will first reach the ear closest to the source and then the ear at the opposite side of the head. At the same time, the ILD refers to the level and spectral differences, induced by the head's filtering effect. The lowest frequency threshold, above which the dominant cue is deemed to be the ILD, is 1500 Hz [27]. At lower frequencies, the wavelength is larger than the ears distance, hence the phase difference is detectable, while at higher frequencies the delay becomes less perceivable. Yet, the threshold is mostly a region of values, because it is related with the distance between the ears, which differs for various listeners. An average value is given to be 22 to 23 cm [12].
However, ITD and ILD aren't necessarily adequate for flawless localization. The front-back confusions occur [25], where the front sounds are perceived as back and vice versa. That happens, when a sound source is located on the surface of the cone and successively yields identical ITD and ILD. That is the so-called cone of confusion problem [28] and it shown in figure 2.2. The phenomenon fades out and the elevation is thoroughly perceived, when the frequency content of the signal is increased, enhancing the ability for monaural localization. Moreover, in [25], it is also pointed out that the increased duration of a stimulus could lead to proper localization as well.

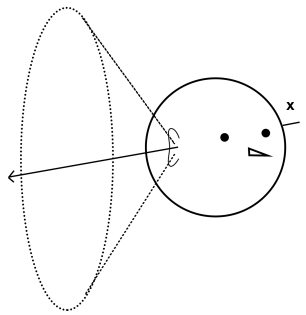


Figure 2.2: Cone of confusion.

In a very reverberant room, the reflections change the original signal that arrives at the ears, causing colouration of certain frequency bands. In that case low-frequency ITD cues are responsible for the correct source localization. The ability of the human hearing system to distinguish the angle of an incidence of a sound wave being firstly radiated by a source, i.e the direct sound component, is called the *precedence effect* [29]. When two sources are present and the delay of the radiation is more than 1 ms, then two separate events are perceived.

2.1.3 Localization of Multiple Sources

The perception of a single sound source is different from that of multiple sound sources. According to Blauert in [25] the sound events can be considered as perceptually separate, when the sound sources are incoherent, whereas coherent sound sources are governed by summing localization. That usually applies for sound sources activated with time difference under 1 ms. Coherent sources activated after a couple milliseconds would be perceived as echo, effect that is audible in rooms with weak absorptive materials. Thus, either identical or signals with difference in level and/or phase, regardless of the frequency content, are considered as coherent.
The multitude in frequency content of several sources being simultaneously active, also constitutes a factor that contributes to perceive different sources as distinct auditory events. For instance, it is remarkable how easily the people can distinguish the instruments' nuance during concert. However, a signal may be masked by another, when the frequency properties are very similar and their level difference is almost identical. In [31], it is proposed that a signal with higher amplitude masks another with lower, within a certain time frame as well as within the same frequency band.
In situations when two loudspeaker signals are coherent or wide-sense coherent, only a single source is perceived. That source is called phantom or virtual and its location mainly depends on the level and phase difference of the signals arriving at the ears. Varying levels of signals in phase is a technique commonly employed in the stereo sound mix. The presence of the phantom source is related to the signal intensity at the loudspeakers and it is explained by the stereophonic law of sines firstly introduced by Blumlein in [32] and depicted in 2.3.

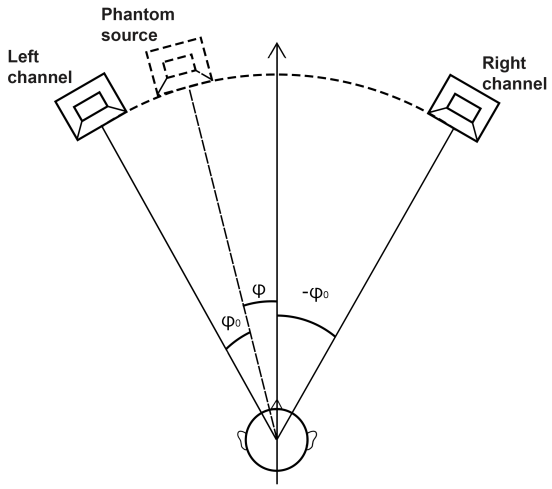


Figure 2.3: Phantom source panning, using a stereophonic reproduction system.

In [10] it is suggested that,

$$\frac{sin\phi}{sin\phi_0} = \frac{a_1 - a_2}{a_1 + a_2} \tag{2.3}$$

where a_1 and a_2 are the gain factors, ϕ is the angle between the axis normal to the head and the direction of the virtual source, and ϕ_0 is the angle between the central axis and the loudspeaker position. For the angles $0°< \phi_0<90°$ and $-\phi_0< \phi<\phi_0$, while the gain take values between 0 and 1.
However, in case the listener's head is rotated, the law of tangent describes more correctly the situation

$$\frac{tan\phi}{tan\phi_0} = \frac{a_1 - a_2}{a_1 + a_2} \tag{2.4}$$

where $0°< \phi_0<90°$ and $-\phi_0< \phi<\phi_0$, while the gain take values between 0 and 1.
It usually happens an auditory event to be created at different angles between the loudspeakers, by controlling the time and/or the level difference of the signals. This effect is attributed to summing localization [32]. Figure 2.3 depicts a single virtual source scenario, appearing close to the left loudspeaker. The for stable loudness of the virtual source, the volume is calculated by

$$C = a_1^2 + a_2^2 \tag{2.5}$$

3

Mathematical Background

3.1 Correlation

The definition of correlation is crucial to the development of the upmix methods presented in the next chapters. It's a measure for description of random data like audio signals. Based on Melchior [33], as autocorrelation function is defined the measure for the linear statistic dependence of two values of a real random variable $x_1(t)$ at a time lag τ:

$$R_{x_1x_1}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_T x_1(t)x_1(t+\tau)dt \tag{3.1}$$

In a similar fashion, the cross-correlation is the measure of a linear statistical correlation of two random variables $x_1(t)$ and $x_2(t)$ at a time lag τ:

$$R_{x_1x_2}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_T x_1(t)x_2(t+\tau)dt \tag{3.2}$$

Then, the cross-correlation coefficient $\Phi_{x_1x_2}(\tau)$ is defined as

$$\Phi_{x_1x_2}(\tau) = \frac{R_{x_1x_2}(\tau)}{\sqrt{R_{x_1x_1}(0)R_{x_2x_2}(0)}} \tag{3.3}$$

To further generalize, when introducing the zero mean functions $x_{1z}(t)$ and $x_{2z}(t)$, equation 3.4 becomes

$$\Phi_{x_1x_2}(\tau) = \frac{\lim_{T\to\infty} \frac{1}{T} \int_T x_{1z}(t)x_{2z}(t+\tau)dt}{\sqrt{\Big(\lim_{T\to\infty} \frac{1}{T} \int_T x_{1z}^2(t)dt\Big)\Big(\lim_{T\to\infty} \frac{1}{T} \int_T x_{2z}(t)^2dt\Big)}} \tag{3.4}$$

where the factors in the denominator represent the quadratic mean and generally defined as

$$\overline{x_z^2(t)} = \lim_{T\to\infty} \frac{1}{T} \int_T x_z(t)^2dt \tag{3.5}$$

3.2 Principal Components Analysis

In general, Principal Components Analysis (PCA) is one of the most powerful statistical methods for multivariate data analysis [34]. It has been used for a variety of

different fields, ranging from natural and social sciences to multiple engineering applications. The aim is to reduce the dimensions of a set of data, without any explicit requirements on the probability density characteristics of the data. The original set of data, including independent variables is linearly transformed to a smaller set of uncorrelated variables, that are easier to understand and manipulate for further analysis. Moreover, as proposed in [35], PCA is also a method for monitoring and identification of correlation patterns among signals. Then, following the notation and description suggested in [36].
Let x_1, x_2, x_3 be three signals and $\mathbf{S}$ is the corresponding cross-spectral density matrix in the frequency domain,

$$\mathbf{S} = \begin{bmatrix} S_{11}(f) & S_{12}(f) & S_{13}(f) \\ S_{21}(f) & S_{22}(f) & S_{23}(f) \\ S_{31}(f) & S_{32}(f) & S_{33}(f) \end{bmatrix} \tag{3.6}$$

which is a Hermitian matrix, that is to say $\mathbf{S} = \mathbf{S}^{*T} = \mathbf{S}^H$, denoting that the matrix is equal to its conjugate transposed. In addition, it is also assumed that the cross-spectral density matrix is written as

$$\mathbf{S}(f) = E[\mathbf{X}^*(f)\mathbf{Y}(f)^T] \tag{3.7}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are column vectors of the same length, ^{T} denotes the corresponding transposed vector and * the complex conjugate, and $E[.]$ is the expectation operator. Suppose that all the rows and the columns of the matrix are linear independent, then $\mathbf{S}$ is written by decomposing the matrix to its eigenvalues:

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^H \tag{3.8}$$

where Λ is the diagonal matrix containing the eigenvalues of $\mathbf{S}$ and $\mathbf{U}$ is the unitary matrix, with rows representing the eigenvectors. To further elaborate, imagine for instance three virtual signals z_1, z_2 and z_3 that are mutually uncorrelated and from which x_1, x_2, x_3 are retrieved. One could also use the more than three virtual signals to make up x_1, x_2, x_3, but for the sake of PCA explanation, lets stick to this concept here. Then, again in the frequency domain, it applies that

$$\mathbf{X}(f) = \mathbf{H}(f)\mathbf{Z}(f) \tag{3.9}$$

or

$$\begin{bmatrix} X_1(f) \\ X_2(f) \\ X_3(f) \end{bmatrix} = \begin{bmatrix} h_{11}(f) & h_{12}(f) & h_{13}(f) \\ h_{21}(f) & h_{22}(f) & h_{23}(f) \\ h_{31}(f) & h_{32}(f) & h_{33}(f) \end{bmatrix} \begin{bmatrix} Z_1(f) \\ Z_2(f) \\ Z_3(f) \end{bmatrix} \tag{3.10}$$

Therefore, the spectral density matrix is defined by

$$\mathbf{S}_{XX}(f) = \mathbf{S} = \mathbf{H}^*(f)S_{ZZ}(f)\mathbf{H}^T(f) \tag{3.11}$$

which is equivalent to equation (3.8), because S_{ZZ} is a diagonal matrix, derived from the uncorrelated z signals, and $\mathbf{H}^* = \mathbf{U}$. It is implied that the eigenvalues of

S are the power spectra of the virtual signals z, which successively means that their magnitudes constitute the principal components.
The virtual coherence function between the ith virtual signal z_i and the jth measured signal x_j is given as

$$\gamma^2_{z_i x_j}(f) = \frac{|S_{z_i x_j}(f)|^2}{S_{z_i z_i}(f) S_{x_j x_j}(f)} \tag{3.12}$$

where the coherence is estimated by the eigenvalues and eigenvectors of **S** and $S_{z_i x_j}(f) = h_{ji}(f) S_{z_i z_i}(f)$.

System Identification with PCA
-See [36]: It will probably help to the presentation.
Suppose **X** denotes a column vector of observations with correlation matrix

$$R_{XX} = E[\mathbf{X}\mathbf{X}^T] \tag{3.13}$$

As discussed in the previous section, let **X** be determined by a set of uncorrelated processes **Z**, through a transformation process denoted by the matrix **T**

$$\mathbf{X} = \mathbf{T}\mathbf{X} \tag{3.14}$$

and then the correlation matrix becomes

$$R_{XX} = \mathbf{T}E[\mathbf{Z}\mathbf{Z}^T]\mathbf{T}^T = \mathbf{T}R_{ZZ}\mathbf{T}^T \tag{3.15}$$

where $R_{ZZ} = \Lambda$ containing the eigenvalues of R_{XX}, because the observations in **Z** are considered uncorrelated. So,

$$R_{XX} = \mathbf{T}\Lambda\mathbf{T}^T \tag{3.16}$$

.
The previous equation express the decomposition of R_{XX} to its eigenvalues and corresponding eigenvectors, which are the columns of the orthogonal matrix **T**

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \tag{3.17}$$

Assuming that x, y are the input and the output of a system and zero mean

$$R_{XX} = E\left[\begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix} \right] = \begin{bmatrix} E[xx] & E[xy] \\ E[xy] & E[yy] \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \tag{3.18}$$

where σ_x^2 is the variance of the variable x, given by

$$\sigma_x^2 = E[(x - \mu_x)^2] \tag{3.19}$$

The eigenvalues are

$$det(R_{XX} - \lambda\mathbf{I}) = \begin{vmatrix} \sigma_x^2 - \lambda & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 - \lambda \end{vmatrix} = 0 \tag{3.20}$$

that gives

$$\lambda_{1,2} = \frac{\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2 + \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2} \tag{3.21}$$

The eigenvectors t1 and t2 corresponding to these eigenvalues are orthogonal and define a basis set of the data, as shown in Figure 11.7.

3.3 Least Squares Oprimization

Again, according to [36], imagine two random variables X and Y, which might be considered having an input(X)-output(Y) connection. Let x_i and y_i be measures of this process, where $i = 1, 2, ..., N$. The aim is to find a linear relationship between the variables, so as $y = bx$. The optimal linear relationship is determined by the slope b. Therefore the error has to be measured between the line and the data points of the set. Two cases are depicted in figure 3.1

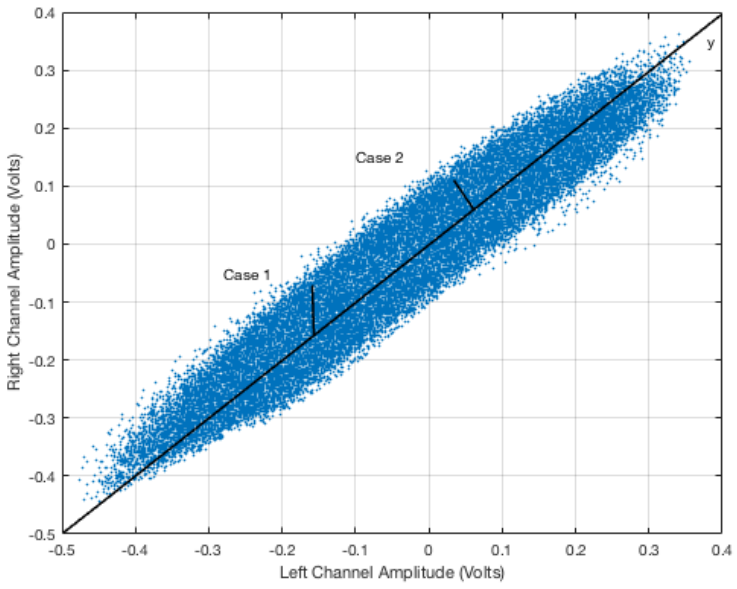


Figure 3.1: The linear relationship of x and y, and two types of error.

In Case 1 the error is calculated in the direction of y axis. Since y is the output of the system, potential errors of the input are neglected. In Case 2, both input and output errors are taken into account, by choosing its direction normal to the output line.

Case 1

The factor b_1 optimizes $y = b_1 x$ and minimizes the sum of the square errors $\sum_{i=1}^{N}(e_i)^2$. The definition of the error is

$$e_i = y_i - b_1 x_i \tag{3.22}$$

The error function or the so-called cost function J is introduced:

$$J = \frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 x_i)^2 \tag{3.23}$$

The error has to be minimized with respect to b_1, thus the minimum of the function needs to be found. The derivative of the cost function considering b_1 as variable is set equal to 0

$$\frac{dJ}{db_1} = \frac{2}{N}\sum_{i=1}^{N}(y_i - b_1 x_i)(-x_i) = 0 \tag{3.24}$$

Then b_1 is computed by

$$b_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{3.25}$$

The numerator is the cross-correlation of two variables x and y and the denominator is the variance of x assuming zero mean. Then, assuming a large number of samples N:

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2} \tag{3.26}$$

Case 2

In contrast to the previous case the error is defined as vertical to the line $y = b_2 x$. The process followed for estimating the factor b_2 is the same as before Then, it is calculated by

$$e_i = \frac{y_i - b_2 x_i}{\sqrt{1 + b_2^2}} \tag{3.27}$$

The cost function becomes

$$J = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - b_2 x_i)^2}{1 + b_2^2} \tag{3.28}$$

Then again, in order to derive the minima

$$\frac{dJ}{db_2} = \frac{1}{N}\sum_{i=1}^{N}\frac{2(y_i - b_2 x_i)(-x_i)}{(1 + b_2^2)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - b_2 x_i)^2(2b_2)}{(1 + b_2^2)^2} \tag{3.29}$$

The result is

$$b_2^2\sigma_{xy} + b_2(\sigma_x^2 - \sigma_y^2) - \sigma_{xy} = 0 \tag{3.30}$$

which yields

$$b_2 = \frac{(\sigma_y^2 - \sigma_x^2) \pm \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}} \tag{3.31}$$

The solution which best fits equation 3.30 or in other words that minimizes the error function J

$$b_2 = \frac{(\sigma_y^2 - \sigma_x^2) + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}} \tag{3.32}$$

3.4 Frequency Domain NLMS Adapive Filter

The main reason for utilizing frequency domain adaptive filters is to deal with long impulse responses. Implementing filters for such signals in time domain reduces the computational efficiency of the algorithm. Hence, the signal in sectioned to blocks in frequency domain to deal with this problem [50].
The following are already stated in Haykin's [50] and in Usher's [45] work.
In general an input signal vector at time l could be written as

$$\mathbf{x}(l) = [x(l), x(l-1), ..., x(l-N+1)]^T \tag{3.33}$$

Then, if k is the block index and L the block length, time l is linked to the k and L by defining

$$l = kL + i \tag{3.34}$$

where $i = 0, 1, ..., L-1$ and $k = 1, 2, ...$
The input data for a block k can be written in the form of a new matrix as

$$\mathbf{x}(k)' = [\mathbf{x}(kL), \mathbf{x}(kL+1), ..., \mathbf{x}(kL+L-1)]^T \tag{3.35}$$

For k block the weight vector is defined as

$$\mathbf{w}(k) = [w_0(k), w_1(k), ..., w_{L-1}(k)]^T \tag{3.36}$$

The convolution of the input signal with the weight sequence is given by

$$y(kL+i) = \sum_{j=0}^{N-1} w_j(k) x(kL+i-j) \tag{3.37}$$

where $i = 0, 1, 2, ..., L-1$
or

$$\mathbf{y}(k)' = [\mathbf{y}(kL), \mathbf{y}(kL+1), ..., \mathbf{y}(kL+L-1)]^T \tag{3.38}$$

The desired and the corresponding error signal will be $d(kL+i)$ and $e(kL+i) = d(kL+i) - y(kL+i)$
Applying Fourier transform, with the method of overlap-save [51], the N tap weights of the filter are padded with an equal number of zeros, i.e $M = 2N$. That means 50% of overlap

$$\mathbf{W}(k) = FFT\Big[\mathbf{w}(k), 0, ..., 0\Big] \tag{3.39}$$

where the amount of zeros is N
Accordingly, in the frequency domain it is

$$\mathbf{X}(k) = diag\Big(FFT\Big[x(kN-N), ..., x(kN-1), x(kN), ..., x(kN+N-1)\Big]\Big) \tag{3.40}$$

which is an M-by-M diagonal matrix, containing two adjacent blocks of samples.
It will also be that

$$\mathbf{Y}(k) = \mathbf{X}(k)\mathbf{W}(k) \tag{3.41}$$

and

$$\mathbf{y}(k) = \mathbf{F}^{-1}\mathbf{Y}(k) \tag{3.42}$$

where $\mathbf{F}^{-1}$ denotes the inverse Fourier transform of the matrix.

Multiplying two vectors in frequency domain corresponds to circular convolution of the vectors in time domain. That yields undesired wraparound aliasing [45], [52] when converting the output back to time domain. Therefore only the last N samples are retrieved because the other N samples are the product of circular convolution.
For the error, it can also be assumed that

$$\mathbf{e}(k) = [e(kN), e(kN+1), ..., e(kN+N-1)]^T \tag{3.43}$$

In the frequency domain the error signal is padded with N zeros as before, in order to avoid aliasing.

$$\mathbf{E}(k) = FFT\Big[0, ..., 0, \mathbf{e}(k)\Big] \tag{3.44}$$

From Haykin [50] the weights in the block LMS algorithm are updated according to the equation

$$\mathbf{w}(k) = \mathbf{w}(k-1) + \mu \sum_{i=0}^{L-1} \mathbf{x}(kL+i)e(kL+i) \tag{3.45}$$

for a *kth* block.
From Parseval's theorem [51], energy density spectrum of the frequency domain block $\mathbf{X}(k)$, given in 3.40, is equal to the power estimate of the signal in time domain. The energy density spectrum is defined as the square of the complex signal. In order to properly normalize the algorithm in equation 3.45 and in frequency domain, a 2N power estimate vector is used according to [45] and [53]:

$$\mathbf{P}(k) = \lambda\mathbf{P}(k-1) + (1-\lambda)||\mathbf{X}(k)||^2 \tag{3.46}$$

where λ is a smoothing constant and $0 < \lambda < 1$.

For a kth block it also applies that

$$\mathbf{m}(k) = \mu \quad diag[\mathbf{P}_0(k)^{-1}, ..., \mathbf{P}_{2N-1}(k)^{-1}] \tag{3.47}$$

where μ is the step-size parameter.
Therefore the normalized equation for the filter update is

$$\mathbf{W}(k) = \mathbf{W}(k-1) + \mathbf{m}(k)\mathbf{X}(k)^H\mathbf{E}(k) \tag{3.48}$$

4

Upmix Methods

The present chapter describes the various upmix techniques belonging to the primary-ambience extraction scheme. They are all implemented in a time-frequency analysis-synthesis manner, using the well studied short time Fourier transform (STFT), except the method of Aarts-Irwan [20], which is developed in time-domain. The rest are initially presented in [17], [22], [24], [13], [19] and they constitute the most important and promising approaches for 2-to-5 conversion. Subjective evaluation experiments have shown that, the listening experiences created with these most of the systems are in general, more preferred than the conventional stereo reproduction.

4.1 Signal Formulation

According to Rumsey in [43], reverberation is added in the stereo recordings, either by artificial or natural fashion. The first refers to the studio case, when audio engineers add digital reverberation in the mix. However, in live recordings, the microphones are usually spatially placed, capturing except the direct signal, the room's sound reflections or the diffuse field of the sound wave. Hence, the ambience is naturally embedded in the recording. Therefore a general signal model is formed in time domain as

$$x_i(t) = \left[\sum_{j=1}^{N} c_j(t) * d_{ij}(t)\right] + \left[\sum_{j=1}^{N} c_j(t) * a_{ij}(t)\right] \tag{4.1}$$

where i is the index for every channel of the stereo signal, i.e. i=1 or 2, $c_j(t)$ represents each source, with N being the total amount of them. These sources are assumed to be convolved with a room impulse response, which include a direct d_{ij} and an ambient component a_{ij}. This convolution process describes the equation (4.1). Yet, Faller follows an alternative signal convention, which will be explained in the corresponding section.
Schroeder in [18], gave a statistical definition for the ambience, that is the exponentially decaying, ergodic and stochastic process normally distributed with a mean of zero. Ergodic means that, the statistical properties of random process can be deduced by a sufficiently long sample derived from the signal.
Since, most methods extract the ambience using STFT, a time-frequency representation of the previous signal model, could be as given by Goodwin-Jot in [22]:

$$\vec{X}_i[k,l] = [x_i[k,l], x_i[k,l-1], ...]^T \tag{4.2}$$

$$\vec{X}_i[k,l] = \vec{P}_i[k,l] + \vec{A}_i[k,l] \tag{4.3}$$

where k is a subband index and l is a time index and each STFT tile is considered as a column vector in time.

Although, the assumptions made are a rational and quite elegant simplification, it is better to be perceived as approximations, since they are not totally solid. However, the methods extracting the ambience in the time-frequency plane, deal also with cases where multiple sources are simultaneously active, thus several direct components exist in the signal. Moreover, in PCA the primary signal has to have higher energy than the ambience, which is also not entirely true in reality.

4.2 PAE with Subjectively Tuned Mapping Function (Method of Avedano Jot)

The method discussed in this section is developed by Avedano-Jot and described in [17]. The ambience levels are deemed to be uncorrelated, having approximately equal energy in every channel:

$$||\vec{A}_i(t)||^2 \simeq ||\vec{A}_j(t)||^2 \tag{4.4}$$

The technique separates the uncorrelated signals in the left and right channels from the direct components, which are correlated in a stereo signal. Then, two new signals are generated including most of the ambient information.
The idea stems from the binaural processing of hearing, which includes the calculation of the cross correlation between the channels, in each critical band [25].
The signal processing front-end consists of a discrete short-time Fourier transform (STFT), converting to the time-frequency plane, in which the correlation at every frequency band will be high in regions where the direct component is dominant, whereas in areas with prevailing reverberation the corresponding correlation will be low.
It has already been mentioned that the STFTs of the channel signals $x_i(t)$ are $\vec{X}_i(k,l)$, where k is the frequency index and l is the time index. Audio signals are in general non stationary, hence the statistics of a signal will change with time. In order to keep track of these changes a forgetting factor λ is introduced. Hence, iterative definition of cross-correlation (i$\neq$ j) and auto-correlation (i = j) is given as:

$$r_{ij}(k,l) = \lambda r_{ij}(k,l-1) + (1-\lambda)\vec{X}_i(k,l)^H \vec{X}_j(k,l) \tag{4.5}$$

where ^{T} indicates transposition, * indicates complex conjugation and ^{H} indicates the corresponding Hermitian matrix. Also, different values of λ can be used in different frequency bands. However, λ is set to a value close to 1 (not 1 though) because the calculation will be ill-defined for $\lambda = 1$ [19].

Then the **inter-channel short-time coherence function** or cross correlation coefficient becomes

$$\Phi(k,l) = \frac{r_{LR}(k,l)}{[r_{LL}(k,l)r_{RR}(k,l)]^{1/2}} \tag{4.6}$$

which is real and will have values close to one in regions where the primary signal is dominant. However, it will be close zero in regions dominated by the reverberation and surrounding noise(ambience).
Extracting the ambience of the stereo, a weighting non-linear function of the short-time coherence is used:

$$\vec{A}_i(k,l) = \vec{X}_i(k,l)\Gamma[\Phi(k,l)] \tag{4.7}$$

where $\vec{A}_i(m,k)$ is the ambience vector having applied Fourier transform and i represents any channel of the stereo. The non-linear function Γ behaves in such a way that low correlated regions are not modified, whereas the high correlated ones are severely attenuated. Hence, the direct component is removed. Furthermore, in order to avoid artifacts, the function has to be smoothed. Since the hyperbolic tangent features this behavior, the mapping function is defined as

$$\Gamma[\Phi] = (\frac{\mu_1 - \mu_0}{2})tanh[\sigma\pi((1-\Phi) - \Phi_0)] + \frac{\mu_1 + \mu_0}{2} \tag{4.8}$$

where μ_1, μ_0 define the range of the output, i.e. the upper and lowest threshold, and σ controls the slope of the function. Figure 4.1 illustrates the masking function (Γ) as function of the cross correlation coefficient Φ and the parameter σ.

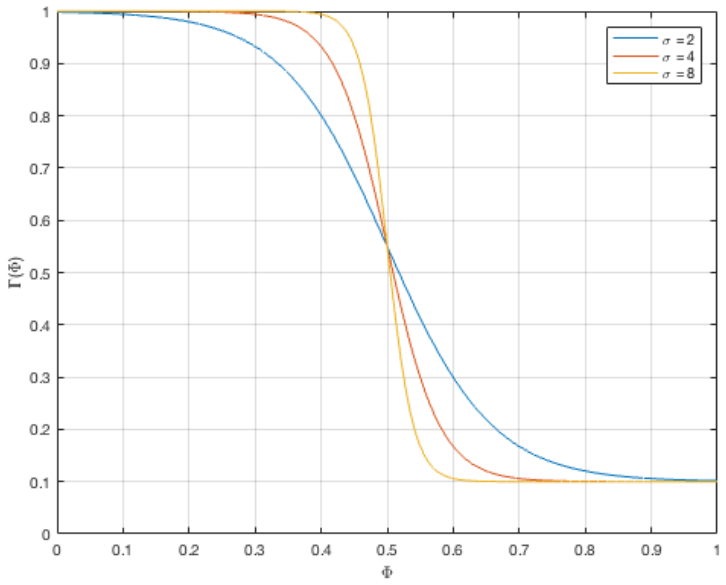


Figure 4.1: Mapping function dependent on the cross correlation coefficient with $\Phi_0 = 0.5$ and different values for parameter σ

In general μ_1 is set to one, since the non-coherent regions has not to be enhanced. On the contrary, μ_0 controls the floor of the function and its value is small, but

greater than zero to minimize subtraction-like artifacts at the output. In the thesis the rest coefficient values are set to $\mu_0 = 0.1$ and $\Phi_0 = 0.5$.
In case the primary signal components are panned completely to one side, the cross-correlation coefficient will be close to zero, that is the extracted ambience is ill-defined. Therefore the signals have to have comparable energies.

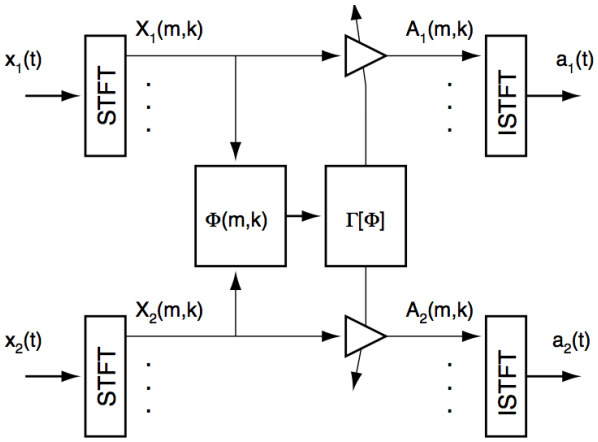


Figure 4.2: Block diagram of ambience extraction, using the mapping function to separate the ambience. It is made according to the diagram presented in [17].

In figure 4.2 a block diagram of the ambience extraction is depicted. The STFT parameters used in every iteration are a Hanning window with of 1024 samples in every iteration, 75% overlap and 256 samples as hop size . Then, the time domain a by the time-frequency equivalentmbience signals are made by applying the ISTFT via the overlap-and-add (OLA) method.

4.3 PAE with Scalar Ambience Masks(Method of Equal Levels of Ambience)

In general, here the auto-correlation and the cross-correlation coefficient are computed as

$$r_{LL} = \vec{X}_L^H \vec{X}_L = \sum_{l=1}^{N} x_L(l)^* x_L(l) = ||\vec{X}_L||^2 \tag{4.9}$$

$$r_{RR} = \vec{X}_R^H \vec{X}_R = \sum_{l=1}^{N} x_R(l)^* x_R(l) = ||\vec{X}_R||^2 \tag{4.10}$$

$$\Phi(k,l) = \frac{r_{LR}}{(r_{LL} r_{RR})^{1/2}} = \frac{\vec{X}_L^H \vec{X}_R}{||\vec{X}_L||\ ||\vec{X}_R||} \tag{4.11}$$

In general the ambience signals are uncorrelated both mutually and with the primary components. Moreover, the cross-correlation coefficient magnitude of the primary signals is one, except for some level and phase differences, which allows for the next assumption

$$r_{LR} = \vec{P}_L^H \vec{P}_R \tag{4.12}$$

and successively

$$|r_{LR}| = ||\vec{P}_L||\ ||\vec{P}_R|| \tag{4.13}$$

As in the subjectively tuned mapping function in section 4.2, the ambience levels in both channels are assumed to be equal, thus

$$I_A = ||\vec{A}_L|| = ||\vec{A}_R|| \tag{4.14}$$

and the ambience levels are calculated according to

$$||\vec{A}_L|| = \alpha_L ||\vec{X}_L|| \quad and \quad ||\vec{A}_R|| = \alpha_R ||\vec{X}_R|| \tag{4.15}$$

These equations imply that the calculation of the ambience energy predicates the determination of the scalar masks α_L and α_R.
For the total signal energy in every channel is reasonable to presume that

$$||\vec{X}_L||^2 = ||\vec{P}_L||^2 + ||\vec{A}_L||^2 \quad and \quad ||\vec{X}_R||^2 = ||\vec{P}_R||^2 + ||\vec{A}_R||^2 \tag{4.16}$$

where $\vec{P}_L$, $\vec{P}_R$ and $\vec{A}_L$, $\vec{A}_R$ are the primary and ambient components respectively. Combining equations (4.13), (4.14) and (4.16) gives

$$|r_{LR}|^2 = I_A^4 - I_A^2(r_{LL} + r_{RR}) + r_{LL}^2 r_{RR}^2 \tag{4.17}$$

It is required that the ambience energy is less than or equal to the total signal energy, i.e. $2I_A^2 \leq r_{LL} + r_{RR}$. Therefore, the solution is

$$I_A^2 = \frac{1}{2}(r_{LL} + r_{RR} - \sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2} \tag{4.18}$$

and as a result from equations (4.19)

$$\alpha_L = \frac{I_A}{|\vec{X}_L||} \quad and \quad \alpha_R = \frac{I_A}{|\vec{X}_R||} \tag{4.19}$$

The ambience energy fraction is estimated as the ratio of the total extracted ambience to the total energy of the inputs, namely

$$E_A = \frac{||\vec{A}_L||^2 + ||\vec{A}_R||^2}{||\vec{X}_L||^2 + ||\vec{X}_R||^2} \tag{4.20}$$

or

$$E_A = 1 - \frac{\sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2}}{r_{LL} + r_{RR}} \tag{4.21}$$

Equation (4.38) is plotted as a function of the level difference between the channels and the cross-correlation coefficient Φ. The graph is shown in figure 4.3 and actually explains the behavior of the method. The ambience energy fraction is 1 only for uncorrelated signals with equal levels, whereas for identical signals with correlation

of 1, the ambience ratio is 0, without being affected by the level difference. When the input signal levels are equal ($r_{LL} = r_{RR}$), the ambience energy fraction is a linear function of Φ. Nonetheless, while the level difference increases, the stronger signal is interpreted as increasingly primary, because the method is based on the assumption that the ambience in both channels has equal levels.

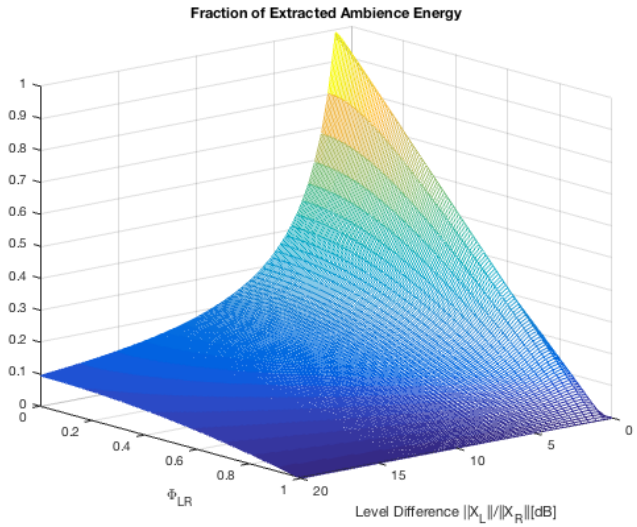


Figure 4.3: Ambience energy fraction as a function of the cross correlation coefficient Φ and the level difference between the inputs

4.4 PAE with Principal Component Analysis (Method of Goodwin-Jot)

The signal model of equation (4.3) is slightly alternated in [22]:

$$\vec{X}_i[k,l] = \vec{P}_i[k,l] + \vec{A}_i[k,l] = \rho_i[k,l]\vec{u}[k,l] + \vec{A}_i[k,l] \tag{4.22}$$

where k is a subband index and l is a time index.
Several assumptions are made as before, in order to adjust the signal properties to the method. However, the notion of orthogonality is explicitly introduced to describe the uncorrelated primary and ambient signals.
The primary component $\vec{u}$ has to be specified, in equation 4.22. This unit vector best describes the set of channel vectors in signal space and it is calculated as the linear combination of the signal vectors

$$\vec{u} = c_L\vec{X}_L + c_R\vec{X}_R \tag{4.23}$$

Then, the orthogonal primary and ambient components can be derived from each channel, by projecting onto u:

$$\vec{P}_i = (\vec{u}^H\vec{X}_i)\vec{u} \tag{4.24}$$

of the ith channel signal and

$$\vec{A}_i = \vec{X}_i - \vec{P}_i \tag{4.25}$$

No signal information is lost during the decomposition because primary and ambient add up to the original. The best choice for the unit vector u lies on the assumption that most of the signal energy corresponds to the primary component, hence it minimizes the energy in the residual ambience. PCA is based on a singular value decomposition (SVD) of the covariance matrix:

$$\left[\vec{X}_L\vec{X}_R\right]\left[\vec{X}_L\vec{X}_R\right]^H = \left[\vec{u}_0\vec{u}_1\right]\begin{bmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{bmatrix}\left[\vec{u}_0\vec{u}_1\right]^H \tag{4.26}$$

where $\vec{u}_0$ and $\vec{u}_1$ are the orthonormal eigenvectors, forming the orthonormal basis for the signal space, and λ_0 is the largest eigenvalue. Therefore, the signals can be written as

$$\vec{X}_L = (\vec{u}_0^H\vec{X}_L)\vec{u}_0 + (\vec{u}_1^H\vec{X}_L)\vec{u}_1 \tag{4.27}$$
$$\vec{X}_R = (\vec{u}_0^H\vec{X}_R)\vec{u}_0 + (\vec{u}_1^H\vec{X}_R)\vec{u}_1 \tag{4.28}$$

Since $\vec{u}_0$ is the unit vector that maximizes the energy, whereas $\vec{u}_1$ minimizes it:

$$E_u = |\vec{u}^H\vec{X}_L|^2 + |\vec{u}^H\vec{X}_R|^2 = \vec{u}^H\left[\vec{X}_L\vec{X}_R\right]\left[\vec{X}_L\vec{X}_R\right]^H\vec{u} \tag{4.29}$$

The maximized and minimized energy are represented by the eigen values by the λ_0 and λ_1 as:

$$E_{u0} = \vec{u}_0^H\left[\vec{X}_L\vec{X}_R\right]\left[\vec{X}_L\vec{X}_R\right]^H\vec{u}_0 = \vec{u}_0^H\lambda_0\vec{u}_0 = \lambda_0 \tag{4.30}$$
$$E_{u1} = \vec{u}_1^H\left[\vec{X}_L\vec{X}_R\right]\left[\vec{X}_L\vec{X}_R\right]^H\vec{u}_1 = \vec{u}_1^H\lambda_1\vec{u}_1 = \lambda_1 \tag{4.31}$$

where the largest and smallest eigenvalues of the matrix XX^H is then computed according to:

$$\lambda_0 = \frac{1}{2}[r_{LL} + r_{RR} + ((r_{LL} - r_{RR})^2 + 4|r_{LR}|^2)^{1/2}] \tag{4.32}$$
$$\lambda_1 = \frac{1}{2}[r_{LL} + r_{RR} - ((r_{LL} - r_{RR})^2 + 4|r_{LR}|^2)^{1/2}] \tag{4.33}$$

which is used to find the vectors $\vec{u}_0$ and $\vec{u}_1$, which occur fro equations (4.27) and (4.28)

$$\vec{u}_0 = r_{LR}\vec{X}_L + (\lambda_0 - r_{LL})\vec{X}_R \tag{4.34}$$
$$\vec{u}_1 = r_{LR}\vec{X}_L + (\lambda_1 - r_{LL})\vec{X}_R \tag{4.35}$$

which in this form are not scaled to the unit-norm. After scaling the primary and ambient component are estimated for each channel as the projection of the channel signal onto the principal vector and minor vectors:

$$\vec{P}_i = \left(\frac{\vec{u}_0^H \vec{X}_i}{\vec{u}_0^H \vec{u}_0}\right) \quad and \quad \vec{A}_i = \left(\frac{\vec{u}_1^H \vec{X}_i}{\vec{u}_1^H \vec{u}_1}\right) \tag{4.36}$$

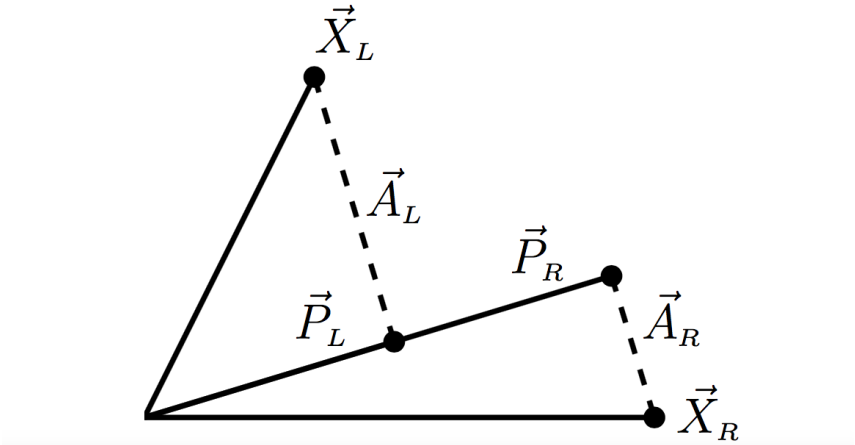


Figure 4.4: Orthogonal decomposition of a stereo signal (X_L, X_R) with principal components analysis (PCA)

The primary and ambience energy fraction derived with PCA can be written as:

$$E_P = \frac{\lambda_0}{r_{LL} + r_{RR}} = \frac{1}{2} + \frac{1}{2}\left(\frac{\sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2}}{r_{LL} + r_{RR}}\right) \tag{4.37}$$

and

$$E_A = \frac{\lambda_1}{r_{LL} + r_{RR}} = \frac{1}{2} - \frac{1}{2}\left(\frac{\sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2}}{r_{LL} + r_{RR}}\right) \tag{4.38}$$

From the previous formulas, it follows that the ambience fraction is 0 for fully correlated signals, whereas the primary is 1, which means that all the signal energy is correctly attributed to primary part of the signal. On the other hand, fully uncorrelated signals of equal energy ($r_{LL} \simeq r_{RR}$) have the same primary and ambient ratio, which is equal to 1/2. The latter implies that, half of the signal energy is ill-defined as primary, a fact which actually stems from the signal assumption that the primary component has most of the signal energy in the mix. As a consequence, PCA behaves more efficiently for inputs which are highly correlated.

4.5 PAE with Least Squares Estimates (Method of Faller)

In the present method, introduced by Faller in [13], direct and diffuse sound are separated with least-squares estimation (multichannel Wiener filtering). In this way, the diffuse signals are statistically independent. A slightly different realization of the signal model and notation is used by the author, which is kept here for fidelity with the method's concept.

In equation 4.39, signal d represents the direct sound from a direction determined by the factor α. The lateral reflections correspond to r_1 and r_2. This, constitutes a decomposition of the stereo with one auditory event (figure 4.5), thus:

$$x_1(n) = d(n) + r_1(n) \tag{4.39}$$

$$x_2(n) = \alpha d(n) + r_2(n) \tag{4.40}$$

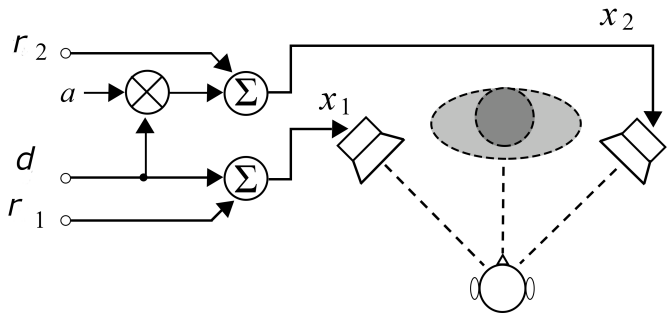


Figure 4.5: Mixing signal s with lateral reflections n_1 n_2. Factor α determines the direction in which the auditory event appears (made according to the equivalent in [13]).

The decomposition is carried out independently in a number of frequency bands in the time-frequency plane, by using the STFT technique-with the parameters as in section 4.2. Hence, the signal model is formulated as:

$$\vec{X}_1(k,l) = \vec{D}(k,l) + \vec{R}_1(k,l) \tag{4.41}$$

$$\vec{X}_2(k,l) = \vec{D}(k,l)\vec{A}(k,l) + \vec{R}_2(k,l) \tag{4.42}$$

where, k is the subband index and l is the time index.
$\vec{X}_1$, $\vec{X}_2$ are stereo subband signals. The aim is to calculate estimates for D, $\vec{A}$, $\vec{R}_1$, $\vec{R}_2$. A short time estimate of the power of X_1 is

$$P_{x_1}(i,k) = E[X_1^2(i,k)] \tag{4.43}$$

where $E[.]$ is the short time averaging operation and $P_R = P_{R_1} = P_{R_2}$.
The normalized cross correlation between left and right:

$$\Phi(i,k) = \frac{E[X_1(i,k)X_2(i,k)]}{\sqrt{E[X_1^2(i,k)]E[X_2^2(i,k)]}} \tag{4.44}$$

which actually can be calculated by the equation (4.6).
A, P_D and P_R are computed as function of the estimated P_{X_1} P_{X_2} and Φ by solving the equation system, where

$$P_{X_1} = P_D + P_R \tag{4.45}$$

$$P_{X_2} = A^2 P_D + P_R \tag{4.46}$$

$$\Phi = \frac{AD}{\sqrt{P_{X_1} P_{X_2}}} \tag{4.47}$$

and they are solved for A, P_D and P_R,

$$A = \frac{B}{2C} \tag{4.48}$$

$$P_D = \frac{2C^2}{B} \tag{4.49}$$

$$P_R = P_{X1} - \frac{2C^2}{B} \tag{4.50}$$

with

$$B = P_{X2} - P_{X1} + \sqrt{(P_{X1} - P_{X2})^2 + 4P_{X2}P_{X1}\Phi^2} \tag{4.51}$$

and

$$C = \Phi\sqrt{P_{X2}P_{X1}} \tag{4.52}$$

Least-Squares Estimation of D, R_1 and R_2

The least squares estimates of D, R_1 and R_2 are functions of A,P_D and P_R.
For each i and k, the signal is estimated as

$$\hat{D} = w_1 X_1 + w_2 X_2 = w_1(D + R_1) + w_2(AD + R_2) \tag{4.53}$$

The estimation error is

$$E = (1 - w_1 - w_2 A)D - w_1 R_1 - w_2 R_2 \tag{4.54}$$

The weights w_1,w_2 are optimal in a LMS sense when the error is orthogonal to X_1 and X_2:

$$E[EX_1] = 0 \tag{4.55}$$

$$E[EX_2] = 0 \tag{4.56}$$

which yields two equations

$$(1 - w_1 - w_2 A)P_D - w_1 P_R = 0 \tag{4.57}$$

$$A(1 - w_1 - w_2 A)P_D - w_2 P_R = 0 \tag{4.58}$$

from which the weights are computed and similarly the R_1 and R_2. Therefore, six weights are estimated, w_1, w_2, w_3, w_4, w_5 and w_6

Postscaling

$\hat{D}$ is scaled to obtain an estimate of D with power P_D

$$\hat{D'} = \frac{\sqrt{P_D}}{\sqrt{(w_1 + aw_2)^2 P_D + (w_1^2 + w_2^2)P_R}} \hat{D} \tag{4.59}$$

$\hat{R}'_1$ $\hat{R}'_2$ are equivalently calculated as

$$\hat{R'}_1 = \frac{\sqrt{P_R}}{\sqrt{(w_1 + aw_2)^2 P_D + (w_1^2 + w_2^2)P_R}} \hat{R}_1 \tag{4.60}$$

$$\hat{R'}_2 = \frac{\sqrt{P_R}}{\sqrt{(w_1 + aw_2)^2 P_D + (w_1^2 + w_2^2)P_R}} \hat{R}_2 \tag{4.61}$$

N'_1 N'_2 which represent the extracted ambience are played with the two side speakers, for stronger impression of envelopment, whereas the estimated A determines the angle σ of the auditory event relative to $\pm\sigma_0$, as in [37] (see also figure 2.3)

$$\sigma = sin^{-1}(\frac{A-1}{A+1} sin\sigma_o) \tag{4.62}$$

Each time-frequency tile the output signal channels are computed as

$$Y_m = \delta(m-1)\hat{N}'_1 + \delta(m-M)\hat{N}'_2 + [\delta(m-l)\alpha_1 + \delta(m-l-1)\alpha_2]\sqrt{1 + A^2 \hat{S}'} \tag{4.63}$$

where m is the output channel index, $2 \leqslant m \leqslant 4$ and $\delta(m) = 1$ at $m = 0$ or 0 otherwise. The signals with 1 and 5 indices are the loudspeakers on the sides. α_1 and α_2 are the amplitude panning factors computed with the stereophonic law of sines.

4.6 PAE with NLMS Adaptive Filter in Frequency Domain (Method of Usher-Benesty)

This following described upmix system is initially presented in [24]. The signals at sample time t are $x_i(t)$ and $x_j(t)$, where $i \neq j$. A summary of the signal processing is shown in figure 4.6. In the proposed ambience extractor upmix system. The filters w_{ij} w_{ji} are adapted over time so that the level of the error signals is minimized.

The purpose of delay D allows for nonminimum phase impulse responses, so as the filtered input to be time-shifted relative to the unfiltered one.

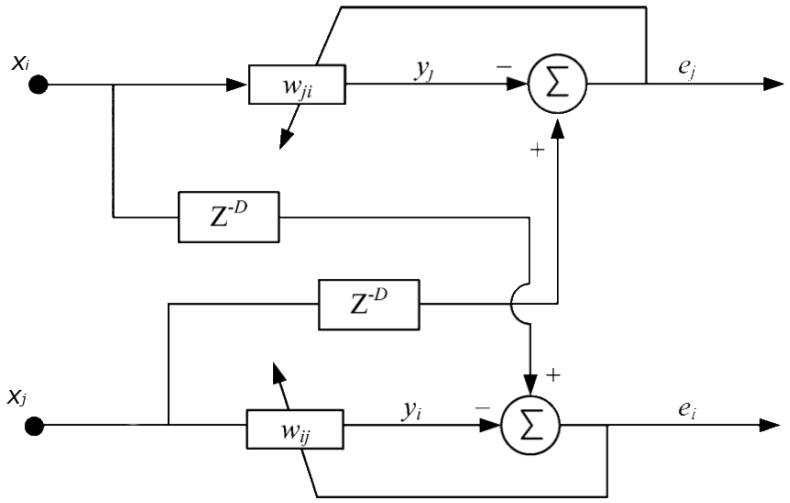


Figure 4.6: Firstly, x_i is filtered with a set of filtering coefficients, i.e. 1024-tap finite-impulse response filter(FIR) and input x_j is time-shifted, typically with a delay of about 10 ms (500 samples). In the next step the difference signal y_j is calculated between the filtered x_i and x_j, which can be transmitted with rear separate loudspeakers. y_j has to be orthogonal to x_i. For this reason , the set of filtering coefficients have to be adjusted accordingly.

Adding a delay to x_i before extracting the y_i allows for time-alignment of the two input signals, in case of the direct sound arrives first in one of the channels. Considering the signals are produced by two microphones spaced up to the 3.4 m, the delay D is set approximately to 10 ms. .

Signal Model

An alternative concept for the signal model is described here. The input $x_i(t)$ (or $x_j(t)$) is defined as

$$x_i(t) = \sum_{l=0}^{L_r-1} s(t-l)d_{i,l} + \sum_{l=L_r}^{L} s(t-l)r_{i,l}, \qquad i = 1 \quad or \quad 2 \tag{4.64}$$

where, the 1st term is the convolution between the source $s(t)$ and the direct part coefficients of L_r length. The 2nd term is the convolution of $s(t)$ with the length reverberant part coefficients of L-L_r length(see figure 4.7).
The time-varying source samples and the time-invariant IRs are defined as vectors:

- $$s_d(t) = [s(t), s(t-1), ..., s(t-L_r+1)]^T \tag{4.65}$$
- $$s_r(t) = [s(t-L_r), s(t-L_r-1), ..., s(t-L)]^T \tag{4.66}$$

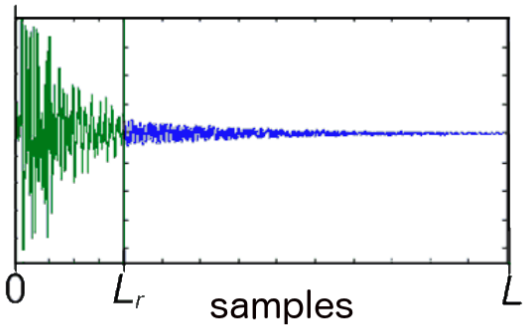


Figure 4.7: Impulse response model for the time-domain transfer function between 2 locations in a room (given in [45]). The first part of the signal occurs up to L_r mixing time and affects the source image. It consists of the primary sound and the early reflections. The rest of the signal tail is the ambience.

- $$d_i = [d_{i,0}, d_{i,1}, ..., d_{i,L_r-1}]^T \tag{4.67}$$
- $$r_i = [r_{i,0}, r_{i,1}, ..., r_{i,L-L_r-1}]^T \tag{4.68}$$

So, equation 4.64 can be written as:

$$m_i(t) = s_d^T(t)d_i + s_r^T(t)r_i \tag{4.69}$$

or, for convenience

$$m_i(t) = s_{d,i}(t) + s_{r,i}(t) \tag{4.70}$$

Definitions and Notation

The ambience extractor must remove the correlated components in the two signals. Therefore, an adaptive filter is applied to one input signal to shape it as similar as possible to the other. Then, by subtracting the filtered signal from the other, the correlated regions are removed.
Filtering x_j by the adaptive filter w_{ij} gives $y_i(n)$:

$$y_i(t) = \sum_{l=0}^{M-1} x_j(n-l)w_{ij,l} \tag{4.71}$$

or in a vector form

$$y_i(t) = {x_j}^T(t)w_{ij} \tag{4.72}$$

where

- $$x_j(t) = [x_j(t), x_j(t-1), ..., x_j(t-M+1)]^T \tag{4.73}$$

•

$$w_{ij} = [w_{ij,0}, w_{ij,1}, ..., w_{ij,M-1}]^T \tag{4.74}$$

The $y_i(n)$ is subtracted from the the unfiltered x_i signal sample-by-sample. Then, the error signal e_i occurs:

$$e_i(t) = x_i(t - D) - y_i(t) \tag{4.75}$$

Optimization Criterion

The algorithm changes the adaptive filter coefficients, thus the aim is to minimize the level of the error signals. This is expressed as a 'performance index' or 'cost' scaler J and it follows that

$$J_i(w_{ij}) = E[e_i(t)^2] \tag{4.76}$$

where $E[.]$ is the statistical expectation operator
When J attains its minimum value, the state of the adaptive filter is called the 'optimal state'.

Adaptation Algorithm

The NLMS algorithm is described in [50]:

$$w_{ij}(t) = w_{ij}(t-1) + \frac{\alpha}{\delta + x_j^T(t)x_j(t)} x_j(t)e_i(t) \tag{4.77}$$

with $0 < \alpha < 1$ and δ is a regularization constant added to the power estimate to ensure against computational problems for low input levels. Implementing the method in frequency domain, 50% overlap is used, as in [45].

Principle of Orthogonality

The optimal state is achieved when

$$E[x_j(t)e_i(t)] = 0 \tag{4.78}$$

A radiated signal by the rear loudspeakers, which is uncorrelated with either x_i or x_j, normally doesn't contain any component affecting the source image.

Assumptions

Using the statistical expectation operator $E[.]$ similar assumptions are made, as in the previous methods:

1. The direct components are at least partially correlated:

$$E[d_i^T(t)d_j(t)] \neq 0 \qquad or \qquad E[s_{d,i}^T(t)s_{d,j}(t)] \neq 0 \tag{4.79}$$

2. The reverberant components are uncorrelated with each other as :

$$E[r_i^T(t)r_j(t)] = 0 \qquad or \qquad E[s_{r,i}^T(t)s_{r,j}(t)] = 0 \tag{4.80}$$

3. The two reverberant IRs are uncorrelated with both early parts as:

$$E[r_i^T(t)d_i(t)] = 0 \qquad or \qquad E[s_{r,i}^T(t)s_{d,i}(t)] = 0 \tag{4.81}$$

4. The two reverberant path IR is decaying random noise with a normal distribution and a mean of zero as follows:

$$E[r_i(t)] = 0 \qquad or \qquad E[s_{r,i}(t)] = 0 \tag{4.82}$$

Selection of the Filter Length

The length of the adaptive filter has to allow for removal of the direct components. Thus, it is considered to be the length of the direct components (L_r). Successively, the reverberant component is defined as the part of an impulse response where the local distribution is normal. It is stated that the ambient samples have kurtosis of 3. Kurtosis is a measure of normality [18]. The definition of kurtosis is

$$kurtosis = \frac{E[x-\mu]^4}{\sigma^4} \tag{4.83}$$

where, μ is the mean and σ the standard deviation of x. In order to determine the start of the reverberant component in the adaptive filter, the kurtosis is calculated for different filter taps, averaging over with 64 samples in every iteration. Taking into account that normal distribution is observed above approximately 1000 samples (23ms), the filter length is set at 1024 samples.

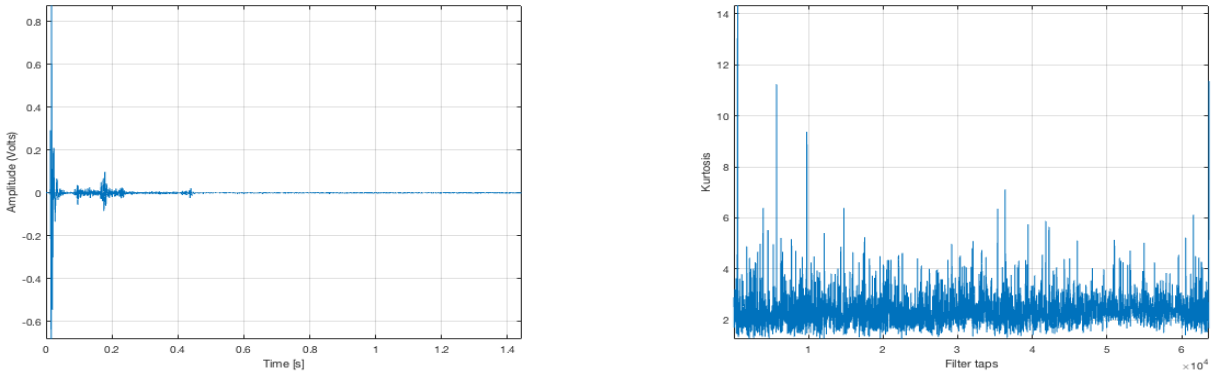


Figure 4.8: Impulse response and the corresponding kurtosis, calculated over 64 samples in every iterration.

4.7 PAE with a Time Domain LMS Filter (Method of Aarts-Irwan)

The algorithm presented here adds a center channel, reformulating the entire front stage and produces a monophonic ambient channel. In order to maintain backward

and forward compatibility, the energy preservation criterion is used to create multi-channel matrices [38].

Center Channel

Klipsch, in [39] suggests that a center loudspeaker improves the quality of stereo sound reproduction, reproducing the signal $\sqrt{2}(x_L + x_R)$. The factor $\sqrt{2}$ preserves the total frontal energy, yet a major drawback is that the image becomes narrow due to crosstalk with L and R.
Based on PCA, the method produces two vectors indicating the direction of the dominant y and the remaining q signal, depicted in figure 4.9. Their directions are perpendicular to each other, formulating a new coordinate system . These are used as basis signals in the matrix decoding
At time index l(integer), each sample of a stereo pair is defined as

$$x(l) = [x_L(l)x_R(l)]^T \tag{4.84}$$

and

$$y(l) = w^T(k)x(l) \tag{4.85}$$

where

$$w(l) = [w_L(l)w_R(l)]^T \tag{4.86}$$

is the weighting vector corresponding to the left and right channels respectively.
The optimal weighting vectors are found by maximizing the energy of equation (4.85) with respect to w:

$$\frac{\partial E[y^2(l)]}{\partial w} = 0 \tag{4.87}$$

where $E[.]$ is the expected value. By means of the steepest descent method [50], a least-mean-square(LMS) algorithm is derived with $y(l-1)$ as input. Hence, for each channel

$$w_L(l) = w_L(l-1) + \mu y(l-1) * [x_L(l-1) - w_L(l-1)y(l-1)] \tag{4.88}$$

$$w_R(l) = w_R(l-1) + \mu y(l-1) * [x_R(l-1) - w_R(l-1)y(l-1)] \tag{4.89}$$

where μ is the step size.
In [40] is shown that the algorithm is stable and input signal dependent if and only if

$$0 < \mu < \frac{2}{x^T(l)x(l)} \tag{4.90}$$

The direction of the stereo image in terms of an angle is given in radians by:

$$\alpha(l) = arctan\left[\frac{w_L(l)}{w_R(l)}\right] \tag{4.91}$$

With L corresponding to $\alpha = \frac{\pi}{2}$ and R to $\alpha = 0$, figure 4.11 shows that α fluctuates around $\pi/4$ creating a phantom source almost equidistant between L and R. The algorithm is able to detect abrupt changes in localization within a short period of time.

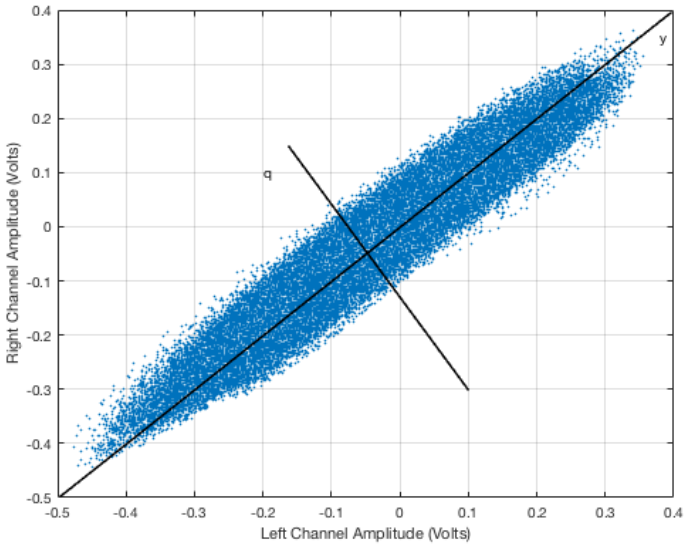


Figure 4.9: Lissajous plot of a highly correlated signal. The bold lines represent the new coordinate system, with y being the dominant signal and q the remaining signal. They form the direction of the stereo image a.

A pair of stereo signals is defined by the unit length vector of equation 4.86. Mapping the stereo vector onto a three-channel vector, the angle α is doubled, creating a new mapping 4.10.b. The projections of the vector are

$$c_{LR} = w_R^2 - w_L^2 \tag{4.92}$$

$$c_C = 2w_L w_R \tag{4.93}$$

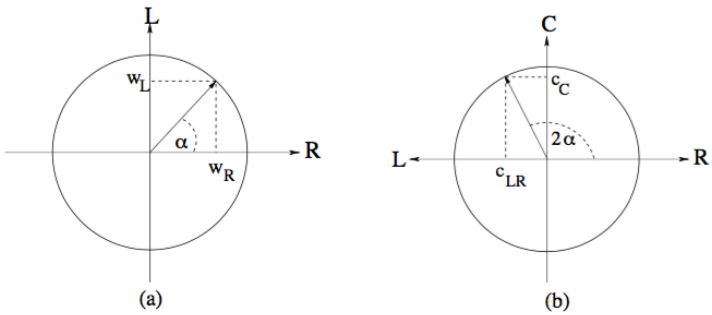


Figure 4.10: (a) Direction of vector plots of stereo signals. (b) Three-channel representation by doubling the angle a.

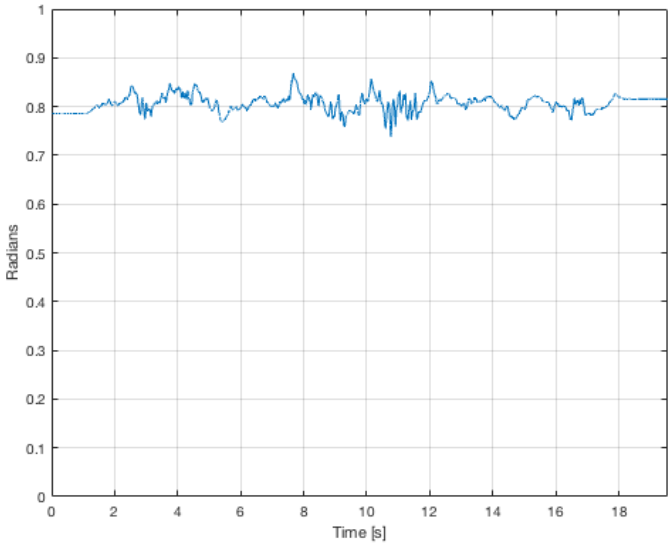


Figure 4.11: Fluctuation of the direction of the stereo image a of the audio excerpt presented in figure 4.9

Surround Channels

The conversion in fig.4.10 works only for non-negative a. If it's negative, no gain can be derived for the central channel. In this case, extra information should be used. The ambience effects can be derived by subtracting the left and right original channels $(x_L - x_R)$, which are represented by the remaining signal q.
In case $|y| \simeq |q|$, the distribution presented in figure 4.9 is no longer an ellipse, but a circle. Hence, a is ill-defined and the correlation coefficient ρ is estimated. Here, the notation is slightly different for the correlation coefficient, because a different iterative calculation procedure is followed, gi in [41]:

$$\hat{\rho} = \hat{\rho}(k-1) + \gamma\Big[2x_L(k)x_R(k) - [x_L(k)^2 + x_R(k^2)]\hat{\rho}(k-1)\Big] \tag{4.94}$$

γ is the step size determining the time constant.
In case $|y| < |q|$, which is possible since $-1 \leq \rho \leq 1$, the input signal are considered uncorrelated:

$$\rho_0 = \begin{cases} \rho, & 0 \leq \rho \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.95}$$

3-D Mapping

Deducing from above, both the direction of the stereo image a and the cross-correlation coefficient ρ are used to avoid ambiguity when $|y| \simeq |q|$. The latter is related with the angle β, shown in figure 4.12, which represents the actual surround information with respect to the front channel sounds:

$$\beta(l) = arcsin[1 - \rho_0(l)] \tag{4.96}$$

and

$$0 \leq \beta(l) \leq \frac{\pi}{2} \tag{4.97}$$

Therefore, it is briefly summarized that,

- q(remaining) increases → weak correlation of inputs → β increases → total distribution to front channels is reduced
- strong correlation of inputs → β approximately zero → larger contribution to the front channels.

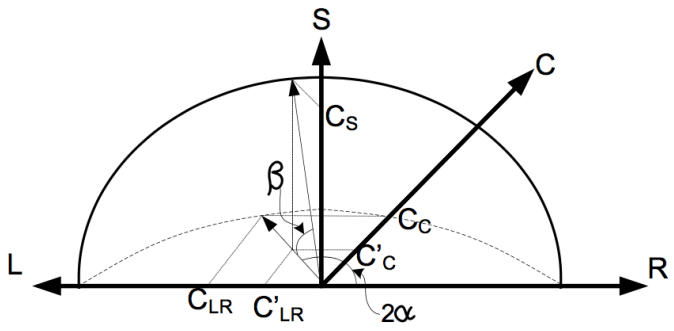


Figure 4.12: 3-D mapping showing front and surround channels.

The recalculation of the projections is necessary, due to the lift of the direction vector (unit length) by an angle β:

$$c'_{LR} = c_{LR} cos\beta \tag{4.98}$$

$$c'_{C} = c_{C} cos\beta \tag{4.99}$$

$$c_{S} = sin\beta \tag{4.100}$$

Matrixing

As mentioned above, the energy preservation criterion has to be fulfilled. A matrix preserves energy if and only if its columns are of unit length and pair-wise orthogonal. If the product of two orthogonal matrices is also orthogonal, the backward-forward compatibility can be achieved. Thus,

$$\begin{bmatrix} u_L(l) \\ u_R(l) \\ u_C(l) \\ u_S(l) \end{bmatrix} = \begin{bmatrix} c_L(l) & gw_L(l) \\ c_R(l) & gw_R(l) \\ c_C(l) & 0 \\ 0 & c_S(l) \end{bmatrix} \begin{bmatrix} y(l) \\ q(l) \end{bmatrix} \tag{4.101}$$

At the left-hand side of eq.4.101 the signals for L, R and C loudspeakers are indicated, while u_S denotes the monphonico surround signal. The basis signals are

obtained by rotating the coordinate system of x_L and x_R, hence a weighted sum of the inputs

$$y(l) = w_L(l)x_L(l) + w_R(l)x_R(l) \tag{4.102}$$

$$q(l) = w_R(l)x_L(l) - w_L(l)x_R(l) \tag{4.103}$$

and

$$c_L = \begin{cases} -c_{LR}, & c_{LR} < 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.104}$$

$$c_R = \begin{cases} c_{LR}, & 0 \leq c_{LR} \\ 0, & \text{otherwise} \end{cases} \tag{4.105}$$

where g is the gain coefficientto control the energy preservation.

5

Up-Mix Results and Discussion

In this chapter, the performance of the algorithms is presented, compared and investigated, based on audio inputs with different correlation characteristics. The results display is divided into two parts. The first discusses the algorithms' possibilities, using a transient pulse and a stereo Gaussian noise signal as inputs. The signals are made in Audacity, containing a correlated primary and an uncorrelated ambient component. They can be perceived as extreme, very specific and simplified cases. The reason for this approach is to provide a fundamental understanding of how the primary-ambience separation works and to efficiently control the study, obtaining a more clear qualitative and quantitative image.

The second focus on three different types of audio signals, each one representing a recording commonly found in a regular musical recording; a single source of congas (see chapter 5.1 for further information about the recording), a regular pop-rock item with multiple sources like drums, guitar, bass and electronics, and an excerpt of electric guitar mostly hard-panned. Their time signals are displayed in figures 5.1, 5.2, 5.3 respectively.

The objective measures, which are used here, aim to quantify the effect the different audio stimuli have on the up-mix results. These measures heve already been developed and introduced in [19]. In the first analysis section, figures are plotted illustrating the ambience energy fraction E_A, ambience level difference L and the cross-correlation of ambience extracted from each channel ϕ_A. Furthermore, since the simple stereo signals are synthesized knowing a priori the primary and ambient components, the primary cross-correlation ϕ_P and the primary-ambient cross-correlation are calculated and presented as well.

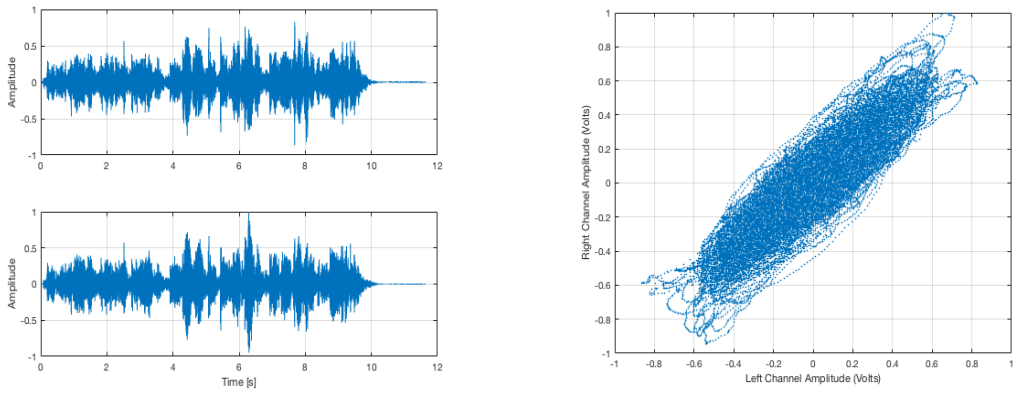


Figure 5.1: Congas stereo channels and their correlation graph. The signal has 0.89 correlation coefficient

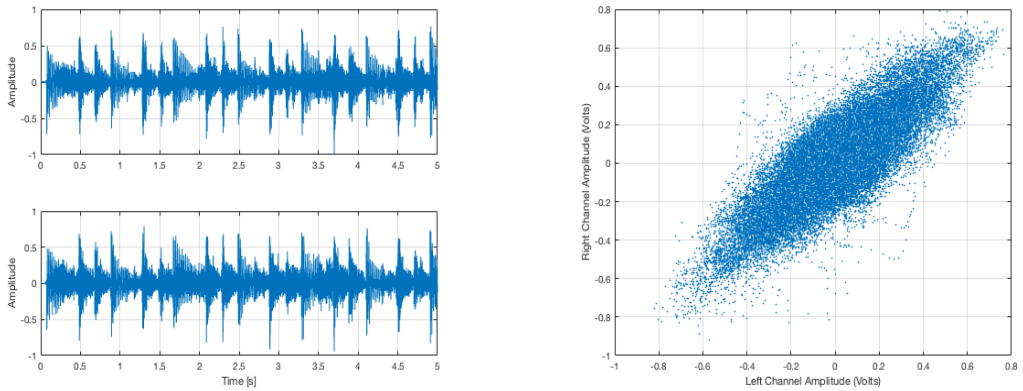


Figure 5.2: Pop-rock stereo channels and their correlation graph. The signals have 0.78 correlation coefficient

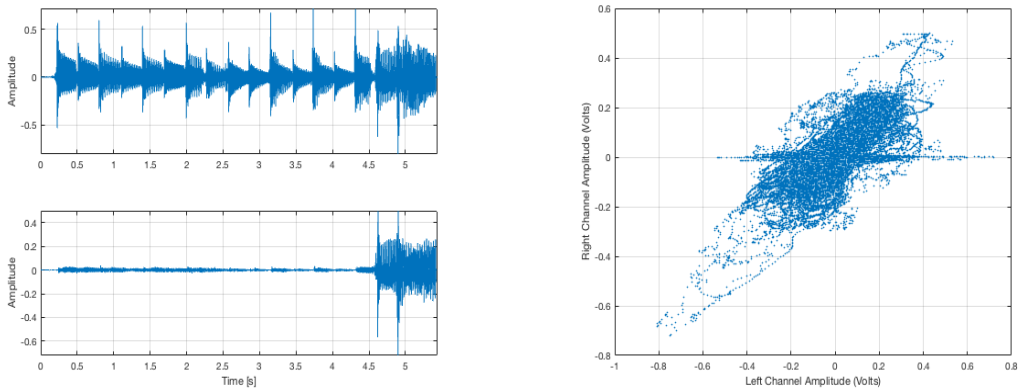


Figure 5.3: Electric Guitar stereo channels and their correlation graph. The signals have 0.48 correlation coefficient

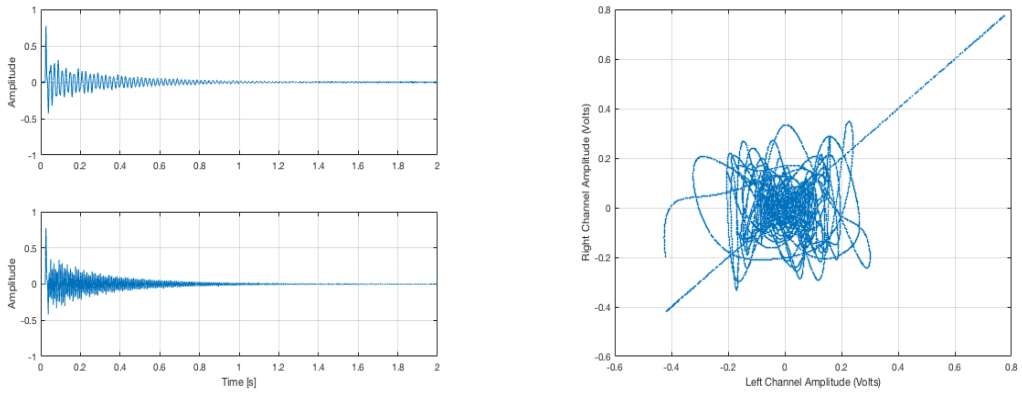


Figure 5.4: On the left, correlated direct signal and uncorrelated decayed ambience in each channel, and on the right the signal's correlation graph with 0.23 correlation coefficient

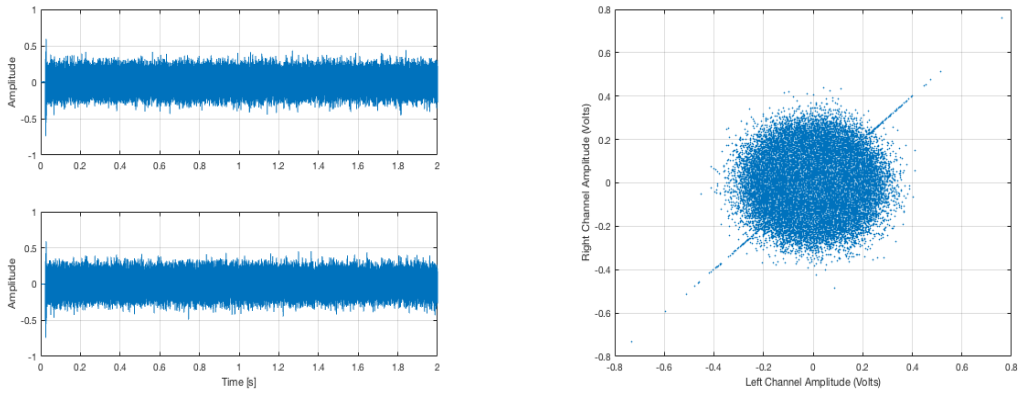


Figure 5.5: On the left, the correlated direct signal and Gaussian uncorrelated noise in each channel, and on the right the signal's correlation graph with 0.014 correlation coefficient

5.1 Evaluation Metrics

The discussed ambience extraction methods have as a mutual property the calculation of the cross-correlation of the left and right input signals. In other words, they attempt to remove the primary components by assuming high correlation of direct components. For this reason, they are assessed within a mutual framework utilizing the same metrics. The metrics have been extracted in every time-frequency frame, with a length of 1024 samples, that is the size of each block of FFT.
At first,the fraction of ambience energy is estimated. E_A is the ratio of the ambience energy extracted from both channels to the total signal energy

$$E_A = \frac{||\vec{A_L}||^2 + ||\vec{A_R}||^2}{||\vec{X_L}||^2 + ||\vec{X_R}||^2} \quad \text{in} \quad dB \tag{5.1}$$

The ambience level difference is defined as the ratio between the left and the right channel. Thus, it can be coalculated by the following formula :

$$L = \frac{||\vec{A_L}||}{||\vec{A_R}||} \quad \text{in} \quad dB \tag{5.2}$$

where $\vec{A_L}$ and $\vec{A_R}$ are the ambient estimated components.
In general, conducting several simulations by using Gaussian Noise the discrepancies are reduced and more secure conclusions are inferred.
Assessing the results in terms of cross-correlation coefficients of the derived ambient and primary components, the ideal values to be expected have to take values between 0 and 1. The formulas for these measures are

$$\phi_A = \frac{\vec{A_L}^H \vec{A_R}}{||\vec{A_L}||\ ||\vec{A_R}||} \tag{5.3}$$

and

$$\phi_P = \frac{\vec{P_L}^H \vec{P_R}}{||\vec{P_L}||\ ||\vec{P_R}||} \tag{5.4}$$

Despite the fact that some methods either extract the energy of the direct signal e.g PCA or produce three frontal channels e.g Aarts-Irwan and Faller's method, the cross-correlation of the primary signal is only used for the simple types of signals (the pulse and the Gaussian noise). The energy of direct sound in every channel is calculated by subtracting the ambience energy $||\vec{A_i}||^2$ from the the total signal energy $||\vec{X_i}||^2$. Therefore, the result of the subtraction yields the autocorrelation of the primary component $||\vec{P_i}||^2$.

5.2 Formulation of Transient Pulse and Gaussian Noise

As mentioned a transient stereo signal is used as input t to test the algorithm. It is shown in figure 5.4. The same sinusoidal pulse in both channels is convolved with different uncorrelated and exponentially decaying artificial room impulse response. Of course, the primary components are identical, implying high correlation. In general the input transient stereo signals are made in accordance with the format of the following the equation.

$$\vec{X_L} = \rho_L \vec{P} + \vec{A_L} \tag{5.5}$$

$$\vec{X_R} = \rho_R \vec{P} + \vec{A_R} \tag{5.6}$$

where $\vec{P}$,$\vec{A_L}$ and $\vec{A_R}$ denote the primary and ambient components, in each channel respectively and ρ_L, ρ_R are the panning coefficients. For the coefficients and the ambient levels, it applies that

$$\rho_R^2 + \rho_L^2 = 1 \tag{5.7}$$

and

$$\|\vec{A_L}\| \simeq \|\vec{A_R}\| \tag{5.8}$$

Gaussian noise signals depicted in figure 5.5 is also used, in order to sufficiently display certain features of the methods, with a deterministic signal as input. In a nutshell, this type of noise is defined as the signal with Gaussian (normal distribution) probability density function (PDF) [54]. This means that, the PDF of a random variable x is normally distributed as:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.9}$$

where μ is the mean value and σ is the standard deviation.

Using such stimuli, it is computationally and optically more efficacious to assess the performance of each algorithm. Let alone, these simple and fixed signals can lead to more secure conclusions, about how the up-mix methods work. In addition, when dealing with more than two system comparisons, the brief duration of these signals is not actually a problem for most of the methods, giving trustworthy results and significantly reducing the simulation processing time.

In the present case the levels of the primary components, existing in both channels, are identical. Hence, the panning coefficients are equal such that $\rho_L = \rho_R$. The ambience is extracted in the frequency domain, under the same STFT scheme, as implemented and mentioned in the previous chapters. A 1024 points analysis and synthesis Hanning window is used, with 75% overlap, to avoid undesired aliasing effects when converting the signals to time domain using the overlap-add synthesis technique.

5.3 Results of Transient Pulse and Gaussian Noise

The following figures show the results obtained for a transient pulse and Gaussian noise, both having correlated direct components. A preliminary examination of the graphs, proves that every method has notable results, separating the previously known embedded ambience.

5.3.1 Mapping Function (Method of Avedano-Jot)

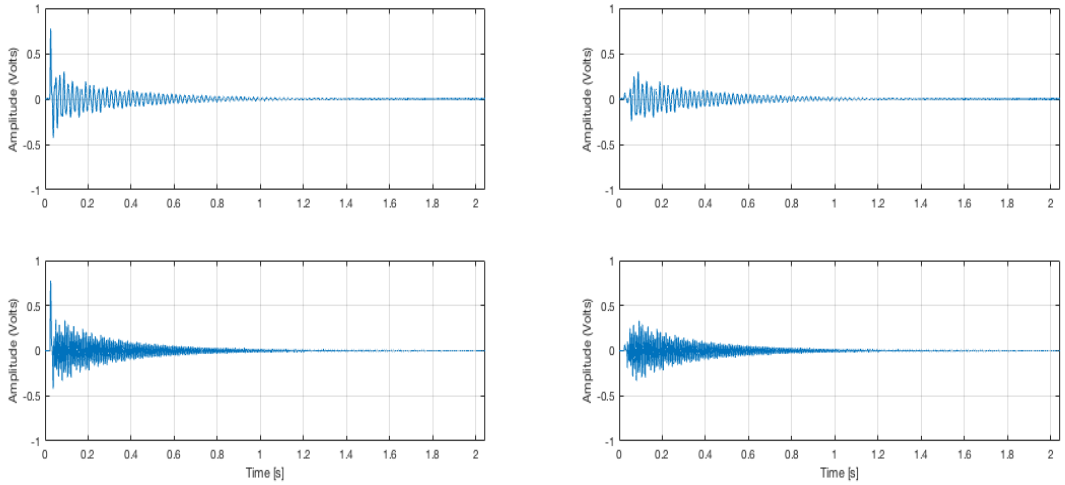


Figure 5.6: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

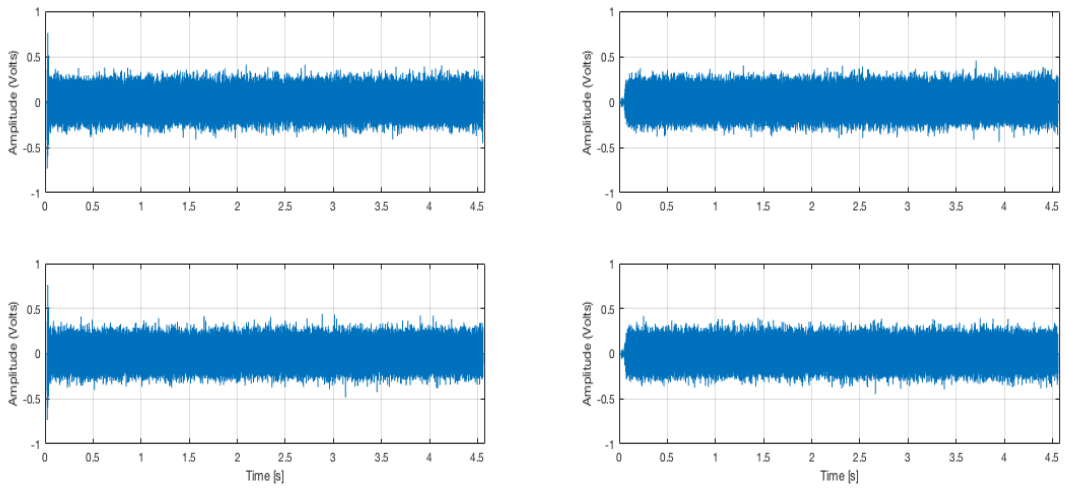


Figure 5.7: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.3.2 Principal Component Analysis

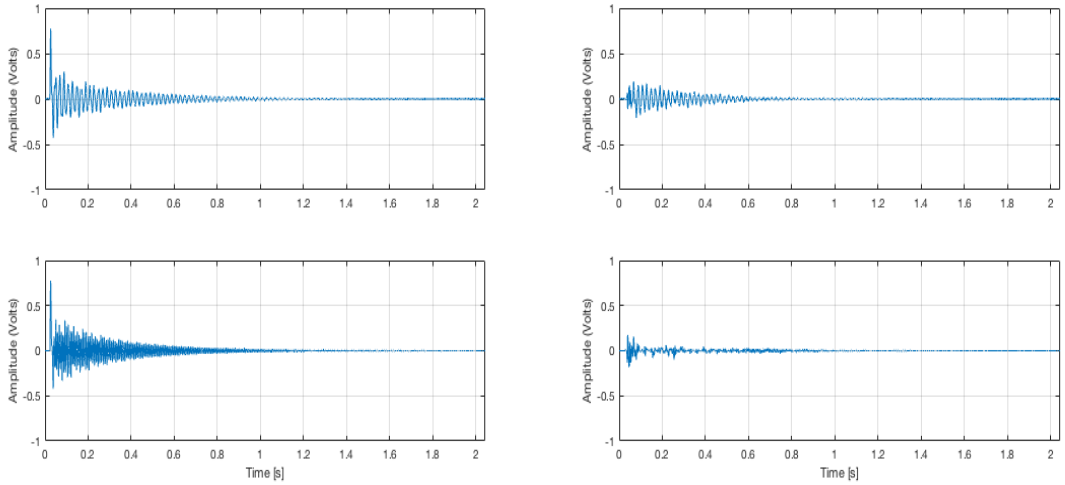


Figure 5.8: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

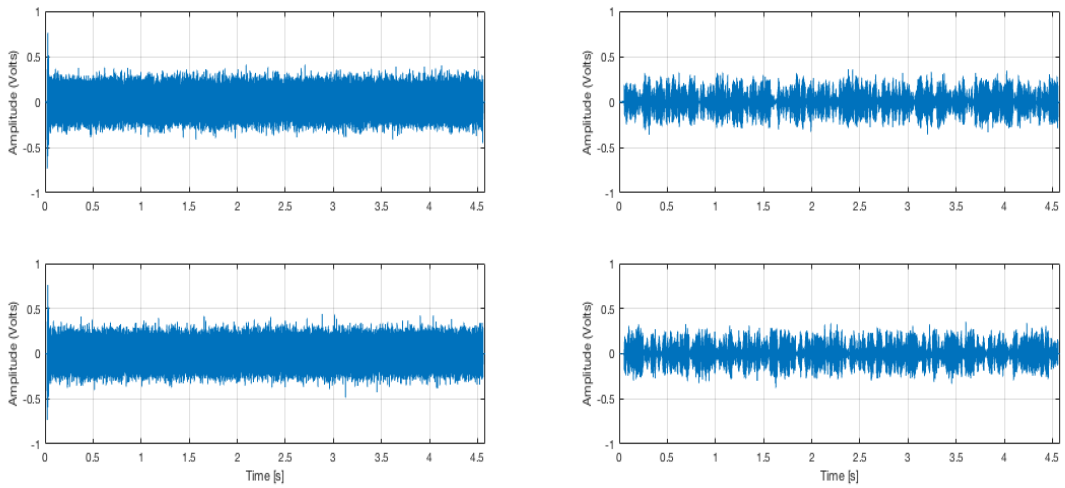


Figure 5.9: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.3.3 Scalar Ambience Extraxtion Mask (Equal Levels of Ambience)

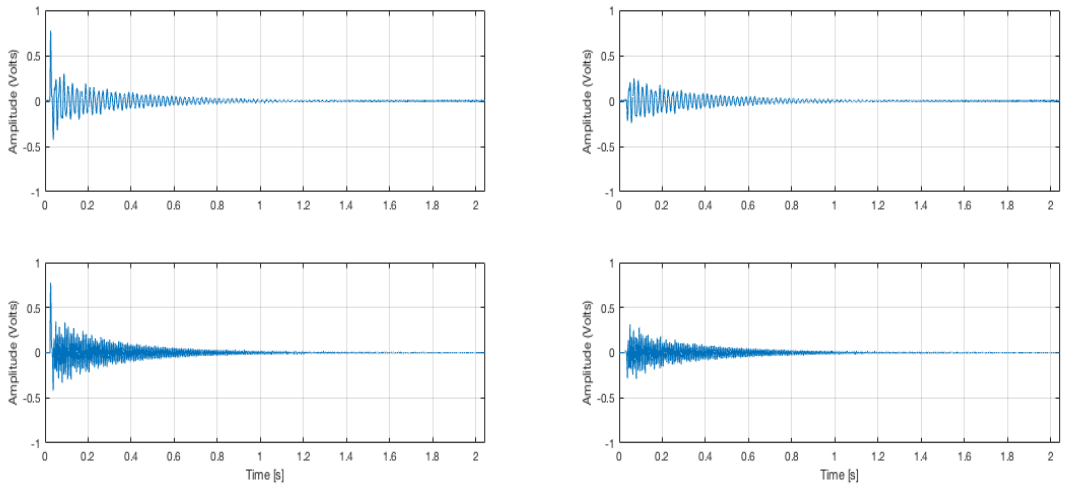


Figure 5.10: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

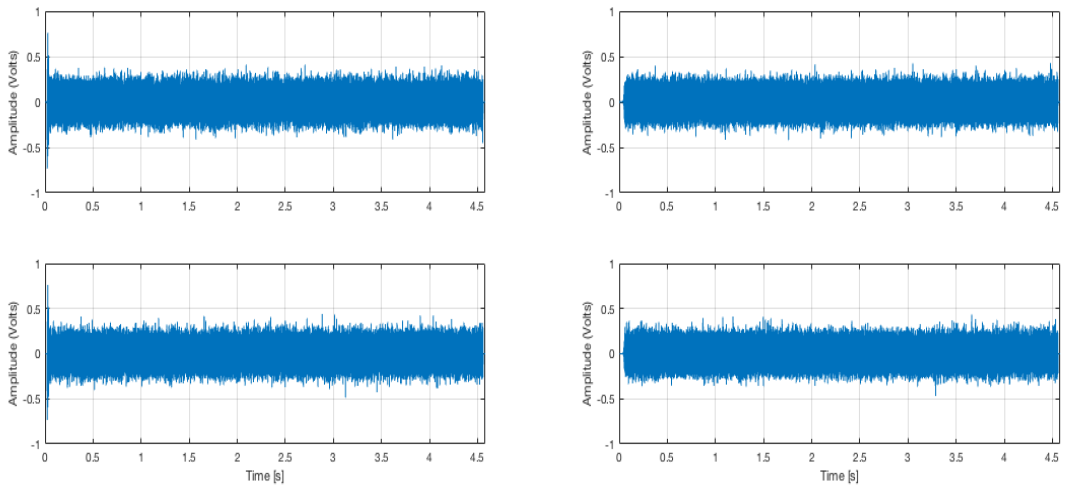


Figure 5.11: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.3.4 Leas Squares Estimates (Method of Faller)

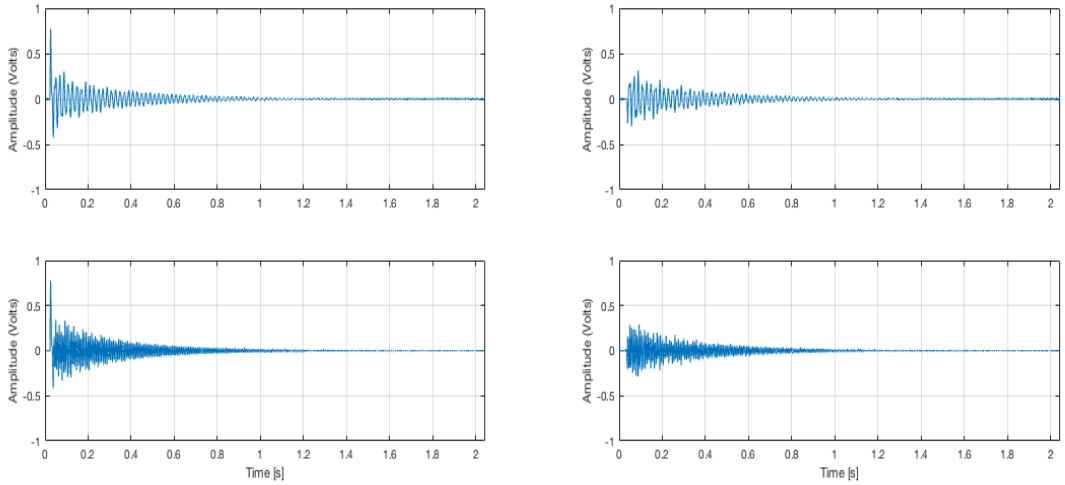


Figure 5.12: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

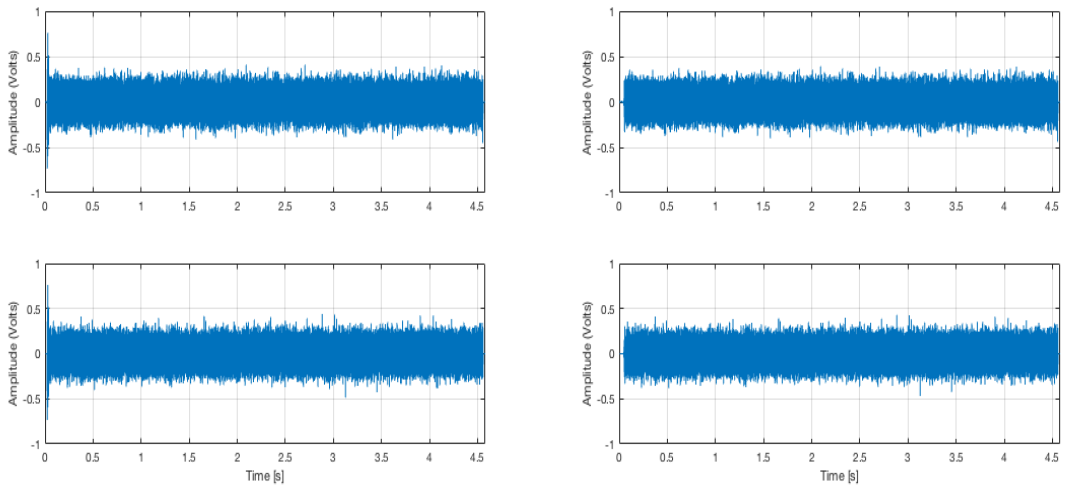


Figure 5.13: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.3.5 Time Domain LMS Filter (Method of Aarts-Irwan)

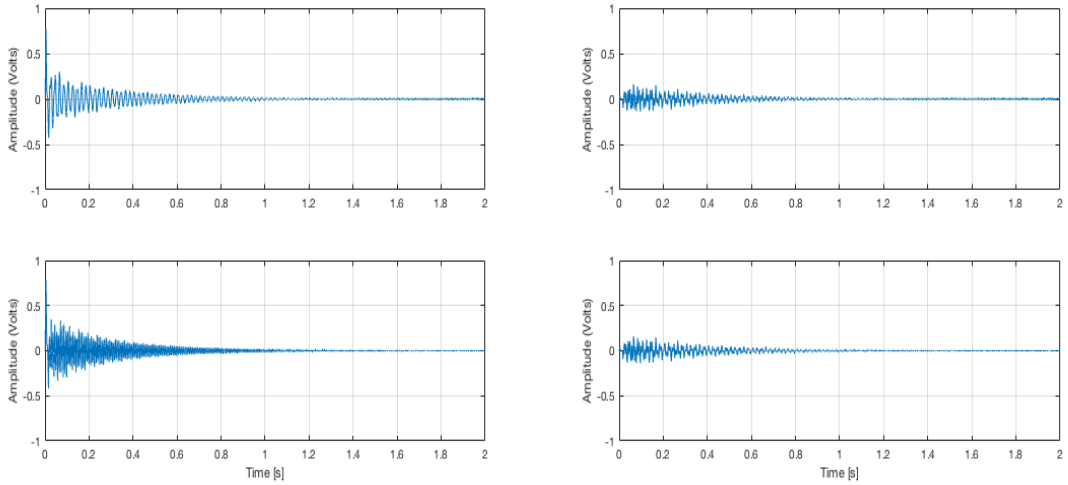


Figure 5.14: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

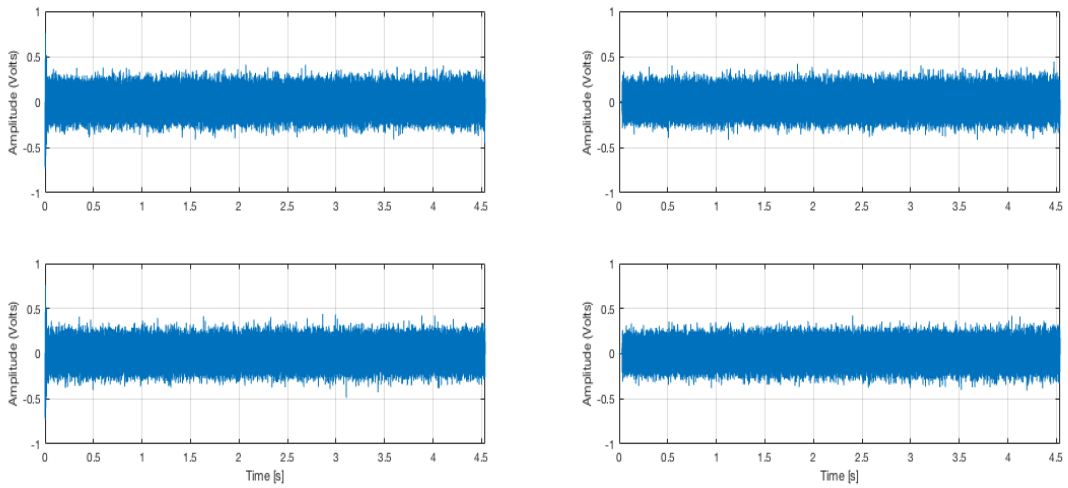


Figure 5.15: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.3.6 Normalized Least Mean Squares (Method of Usher-Benesty)

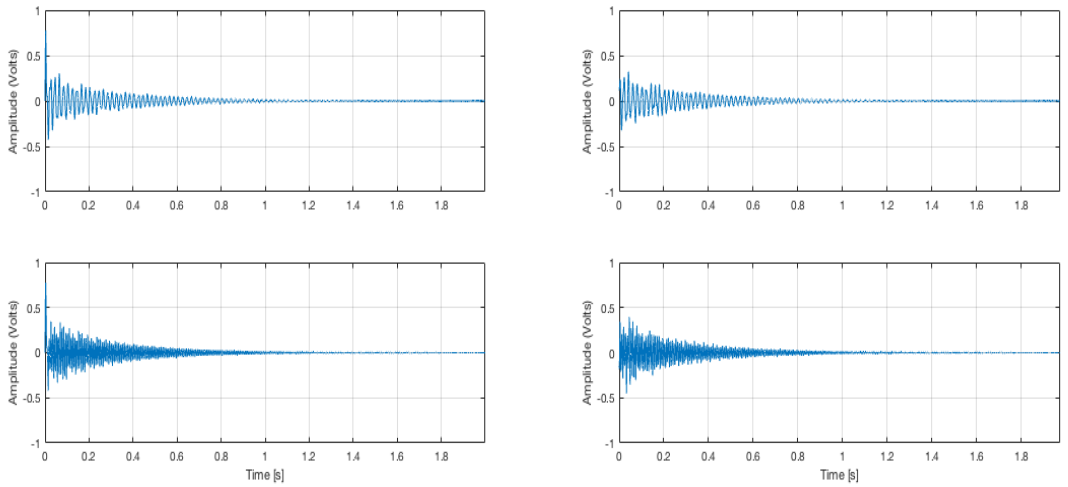


Figure 5.16: Initial stereo pulse with correlated direct components, convolved with room impulse responses, and the corresponding extracted ambience.

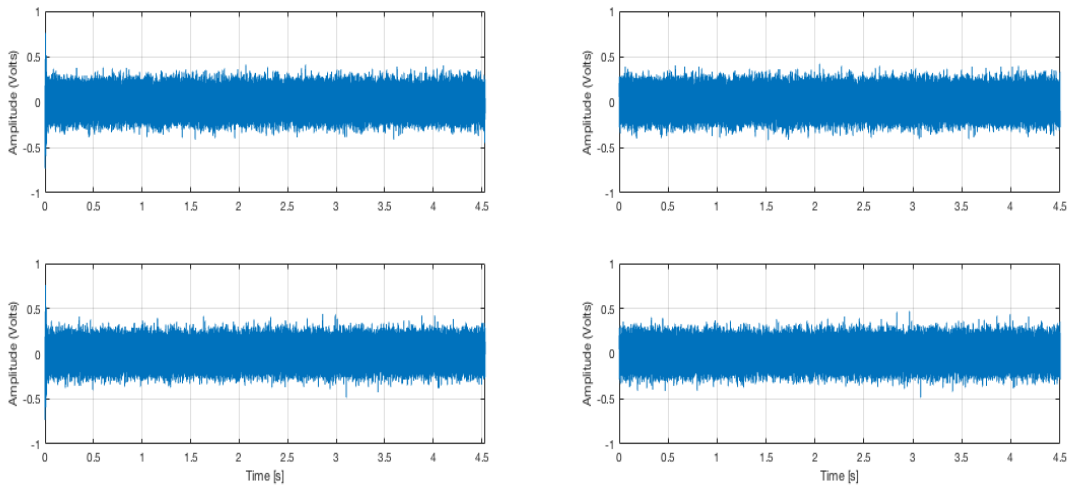
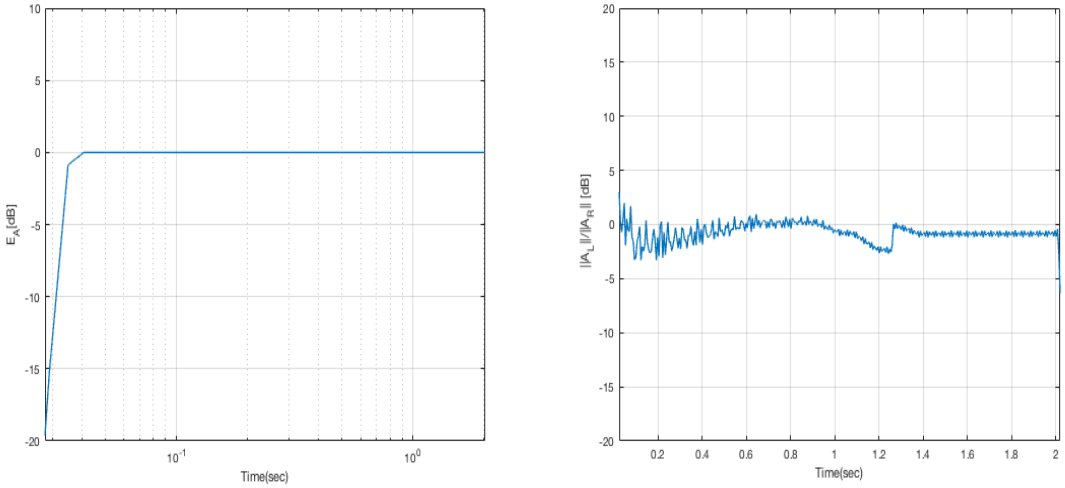


Figure 5.17: Initial Gaussian noise with correlated direct components and the corresponding extracted ambience.

5.4 Evaluation Metrics for the Transient Pulse and Gaussian Noise

The signals depicted in figures 5.4 and 5.5 are used as inputs. In that case there is a prior knowledge for the signal features, which are also made according to the specifications described in section 5.2. Taking into account the dependency of the algorithms' efficiency on the signal properties, one has to better monitor the input samples to thoroughly assess the methods with the recommended measures of section 5.1.

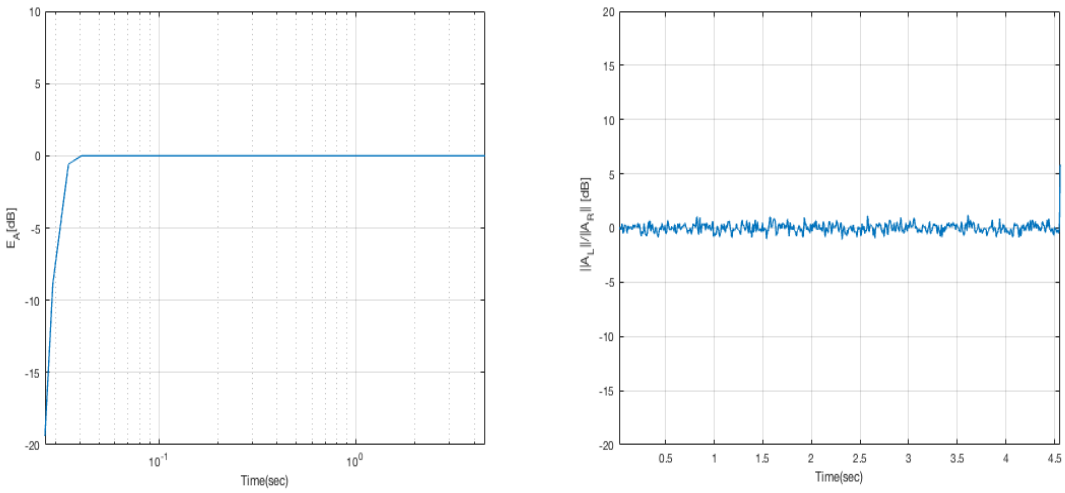
Figure 5.18 shows the ambience energy fraction and the ambience level difference initially produced for the transient signal (shown in figure 5.4). Ideally, the results from the up-mix systems follow the behaviour of the curves depicted in figure 5.18. Furthemore, in figure 5.18b, the level difference fluctuates around or just below 0 dB, indicating a discrepancy from the requirement for equal levels in both channels. This is something to be expected, since the signal is made by a simple convolution between a sinusoidal pulse and an artificial room impulse response, which differs in both channels.



(a) **(b)**

Figure 5.18: Ambient energy fraction and ambient level difference, originally included in the signal, for the transient pulse

The original embedded ambience and level difference of fully uncorrelated Gaussian noise, with correlated direct components, (presented in figure 5.5), are shown in figure 5.19. In this occasion, the ambient levels in the channels are more balanced than before, wiggling steadily with time from -1 to 1 dB.



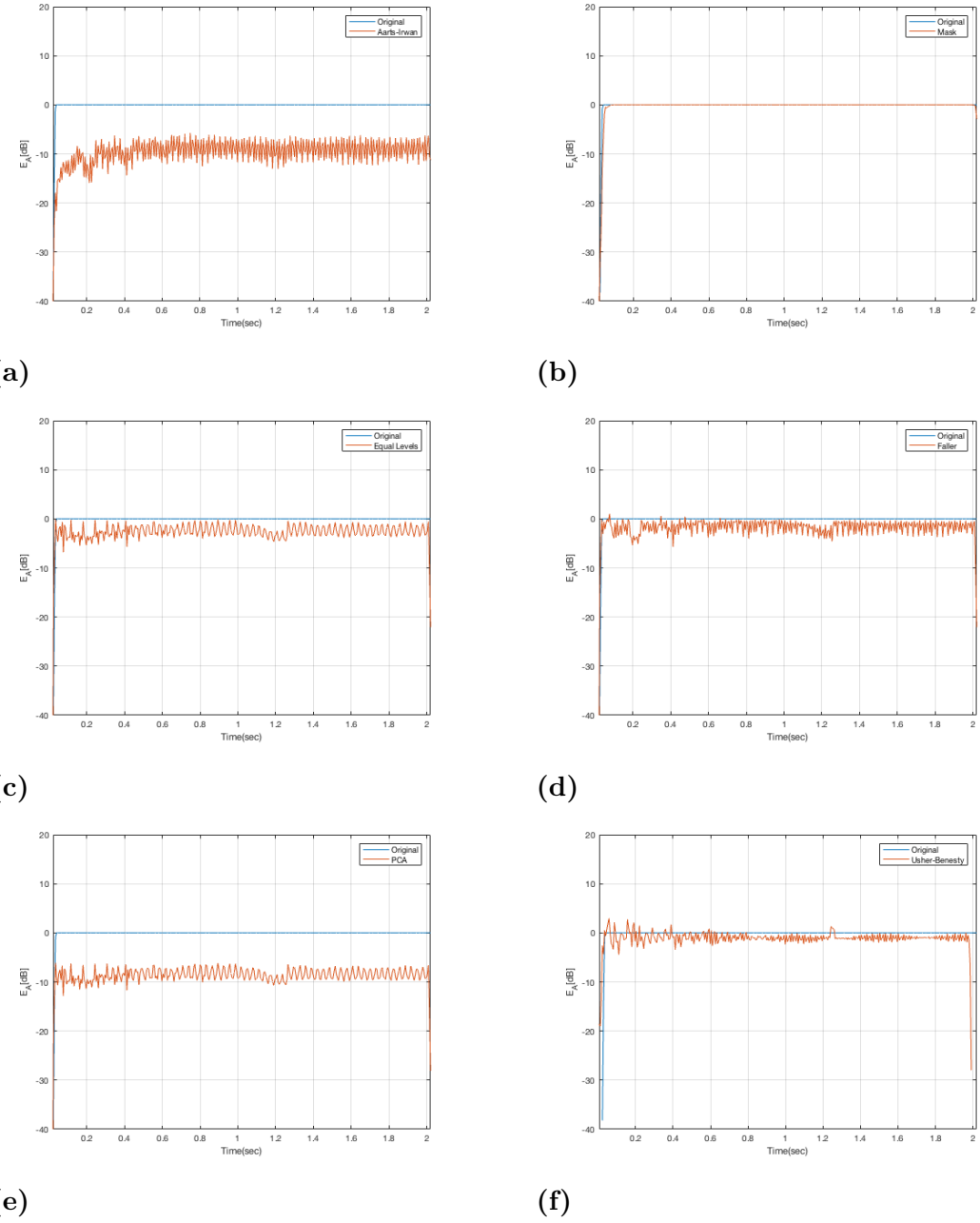
(a) **(b)**

Figure 5.19: Ambient energy fraction and ambient level difference, originally included in the Gaussian noise signal.

5.4.1 Ambience Energy Fraction, E_A

The graphs 5.20 and 5.21 show the ambience energy fraction (E_A) with time, given by equation (5.1). In general, the curves appear to follow the trend of the original embedded fraction. The mapping function among the presented methods achieves the best results, by totally coinciding with the original curve. There is almost no leakage to the primary signal, since it fully coincides with the reference curve of the original contained ambience. Only a small fraction of residual ambient energy is present in the primary component, as it starts increasing slightly later than the reference line [17]. In general, the extracted energy fraction is accurately calculated from most of the methods, except Aarts-Irwan system with the transient as input (figure 5.20a) and PCA. In the first case, the level curve fluctuates between -8 and -12 dB lower than the expected vales. As it is noted in the description of the method in section 4.7, the weights w_L and w_R are calculated by the LMS algorithm and they actually constitute the solutions of the Wiener-Hopf equations (in [50], page 104). Therefore, the step-size μ prevents from reaching an optimal solution and the algorithm performs a random motion around the minimum point, which is the so called gradient noise.

5.4.1.1 Transient Pulse



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.20: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

The pulse is non-stationary, meaning that the statistical properties of the signal vary over time [49] and the filter solutions will be non-stationary as well. Hence, there will be a difference between the optimal and the existent filter state. Notwithstanding, for the Gaussian noise, which is a wide sense stationary signal, namely its mean and autocorrelation properties are time invariant, LMS extracts the ambience with even more precision, as it is depicted in figure 5.21a. The same could apply for the NMLS

algorithm, since it is simply normalized by a regularization constant. However, the E_A for the NMLS (Usher-Benesty) is even closer to the optimal solution for each signal type, yielding a very accurate ambience fraction.

5.4.1.2 Gaussian Noise

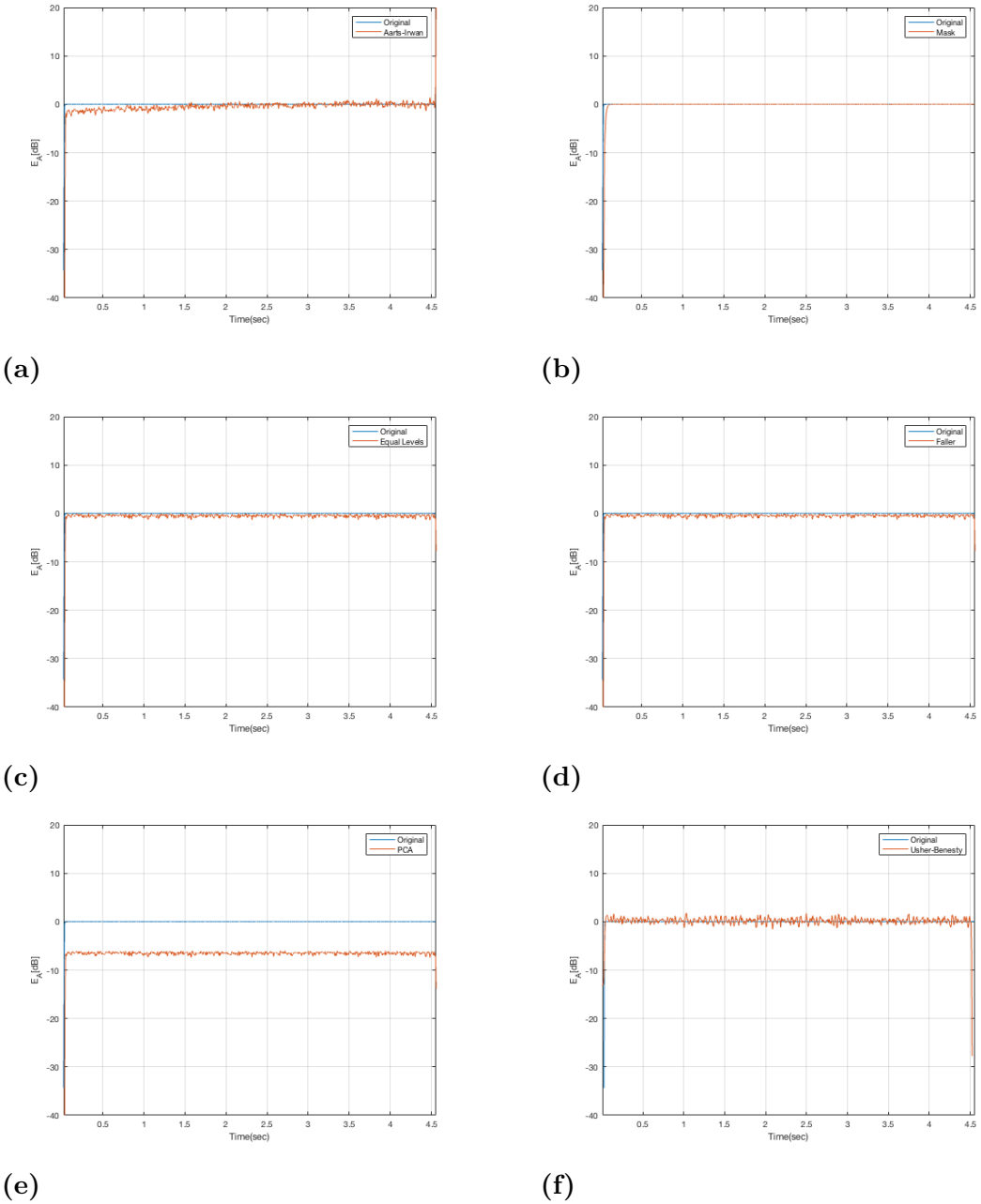


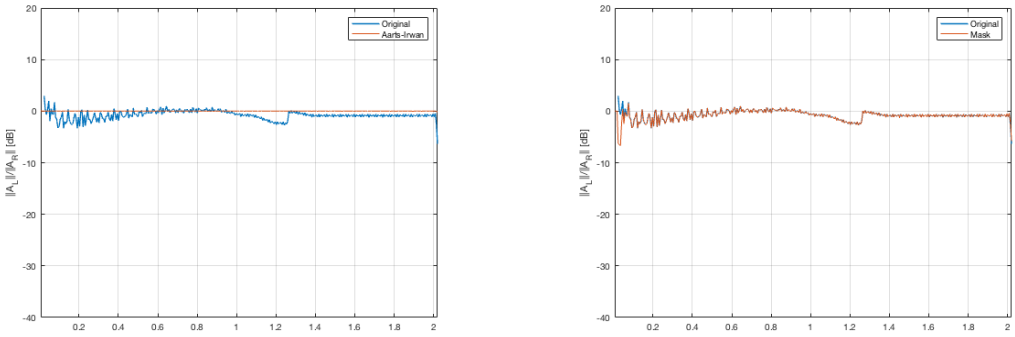
Figure 5.21: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

Equal Levels, PCA and Faller's method appear to have similar behaviour with each other when calculating the energy fraction. They fluctuate between 0 and -5 dB,

with PCA being a bit lower in levels. Therefore, it can be deduced that small portions of ambience signal are misinterpreted as primary. The effect is most prominent in PCA, which is actually in line with the description of the PCA performance in chapter 4.4 and initially defined in [22]. This is an inherent disadvantage of the method, considering the emerging limitation . Signals of similar energy($||\vec{X}_L||^2 = ||\vec{X}_R||^2$) and completely uncorrelated with each other imply that, just about half of the signal energy is mischaracterized as primary. This can be seen in equation (4.38). More precisely, one of the channels is then educed as primary, which subsequently means that PCA becomes less advantageous for stereo uncorrelated channels. It also indicates that certain input signal features are required to obtain reliable results with PCA.

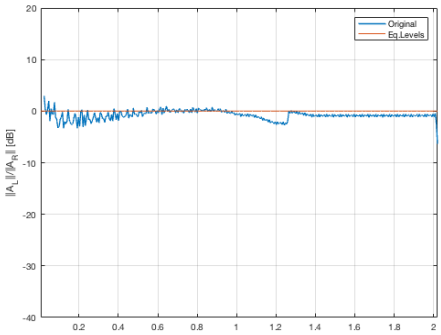
5.4.2 Level Difference, L

5.4.2.1 Transient Pulse

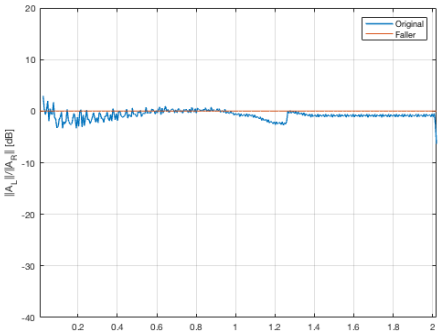


(a)

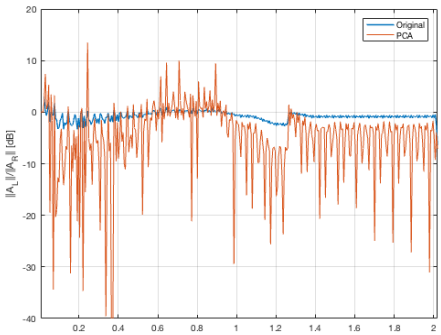
(b)



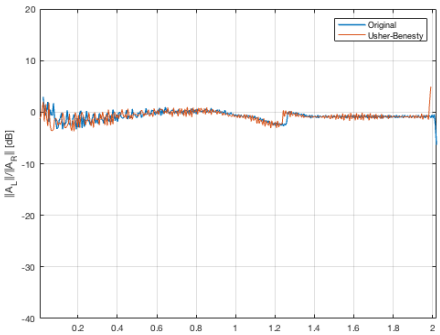
(c)



(d)



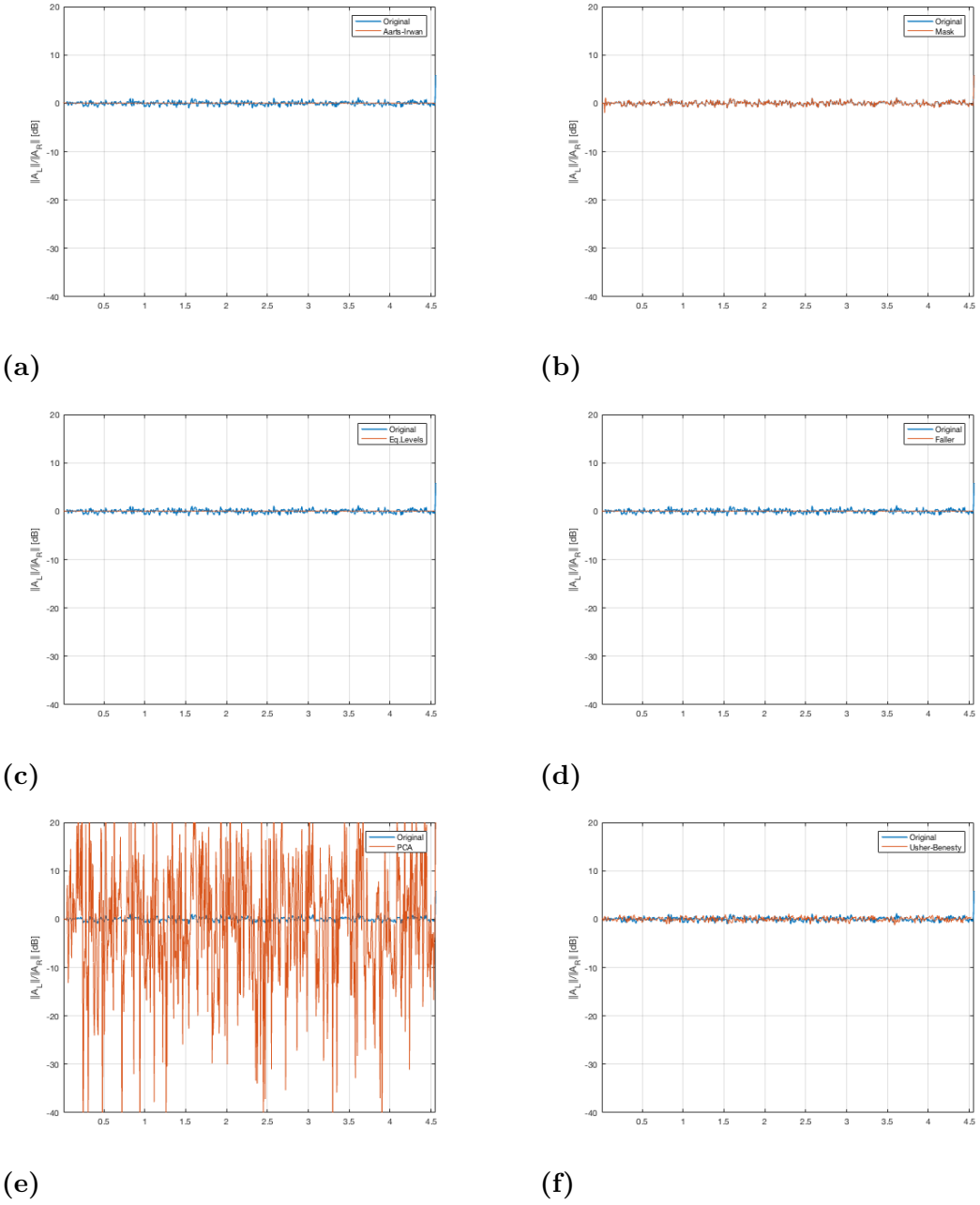
(e)



(f)

Figure 5.22: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.4.2.2 Gaussian Noise



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

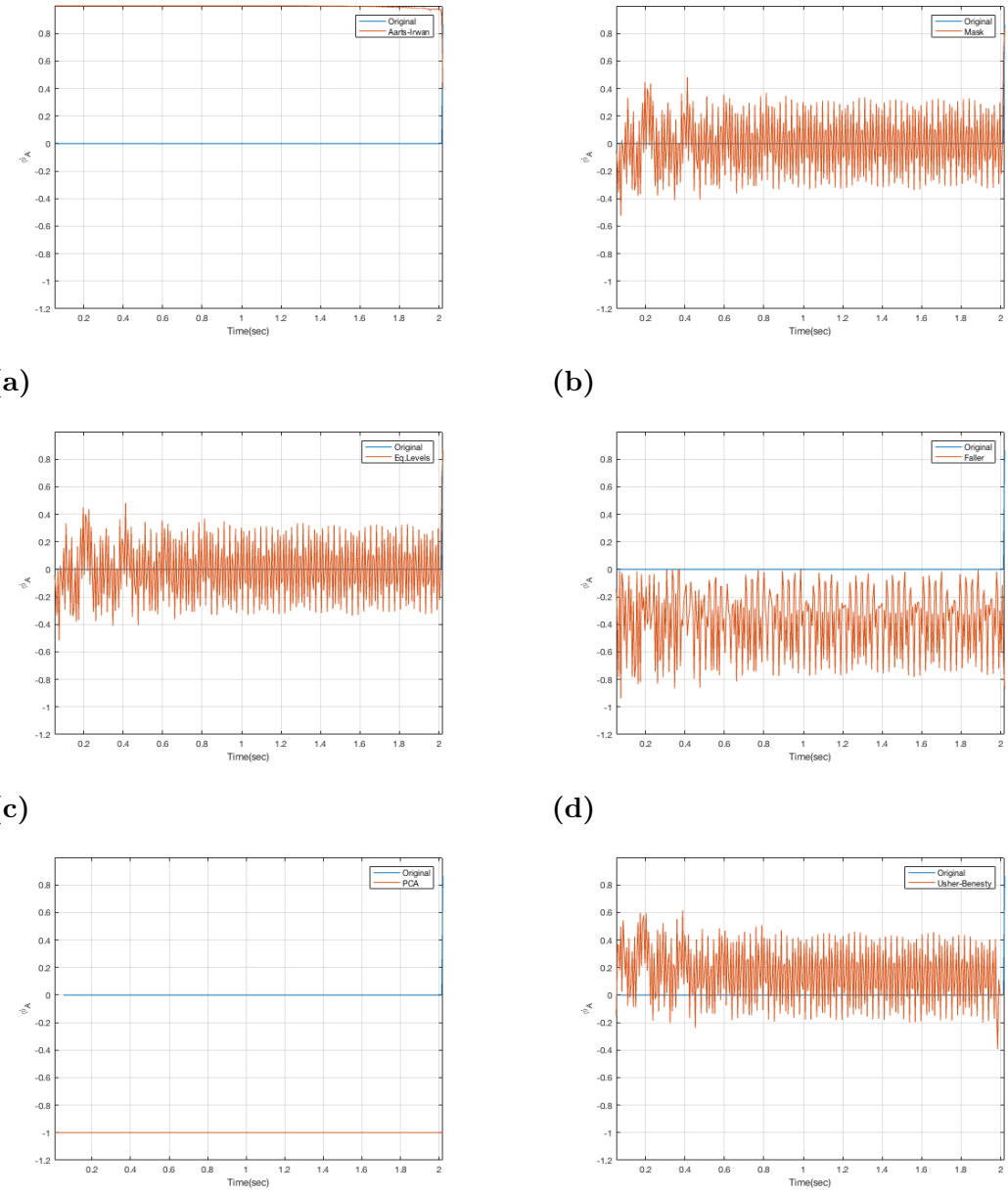
Figure 5.23: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

It is apparent, in regard to the ambient level difference depicted in figures 5.22 and 5.23, that PCA occurs severe fluctuations in the range from 20 to -50 dB. On one hand, the extracted ambience is almost hard-panned to the right channel, which means that almost the entire ambience energy is attributed to the right channel, inducing a quite unfavourable artifact. On the other hand, figure 5.23 shows that PCA curve harshly fluctuates. Accordingly, the ambience is ill-defined in this case

and probably will occur severe artifacts in real-time listening conditions. On the contrary, the mapping function follows, with high precision, the originally isolated ambience, while Equal Levels and Faller's method remove exactly the same amount of ambience from each channel.

5.4.3 Ambience Cross-Correlation Coefficient, ϕ_A

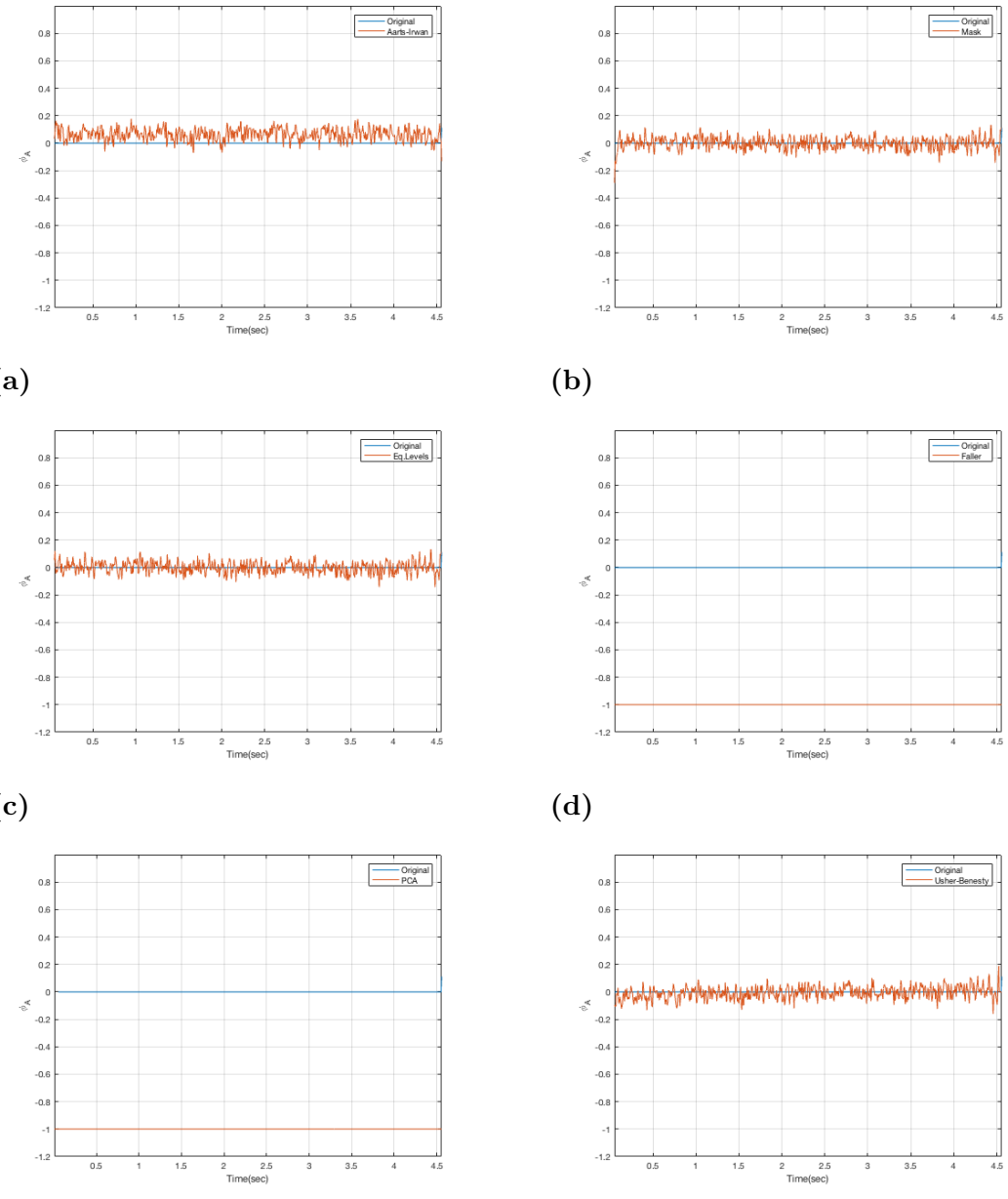
5.4.3.1 Transient Pulse



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.24: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.4.3.2 Gaussian Noise

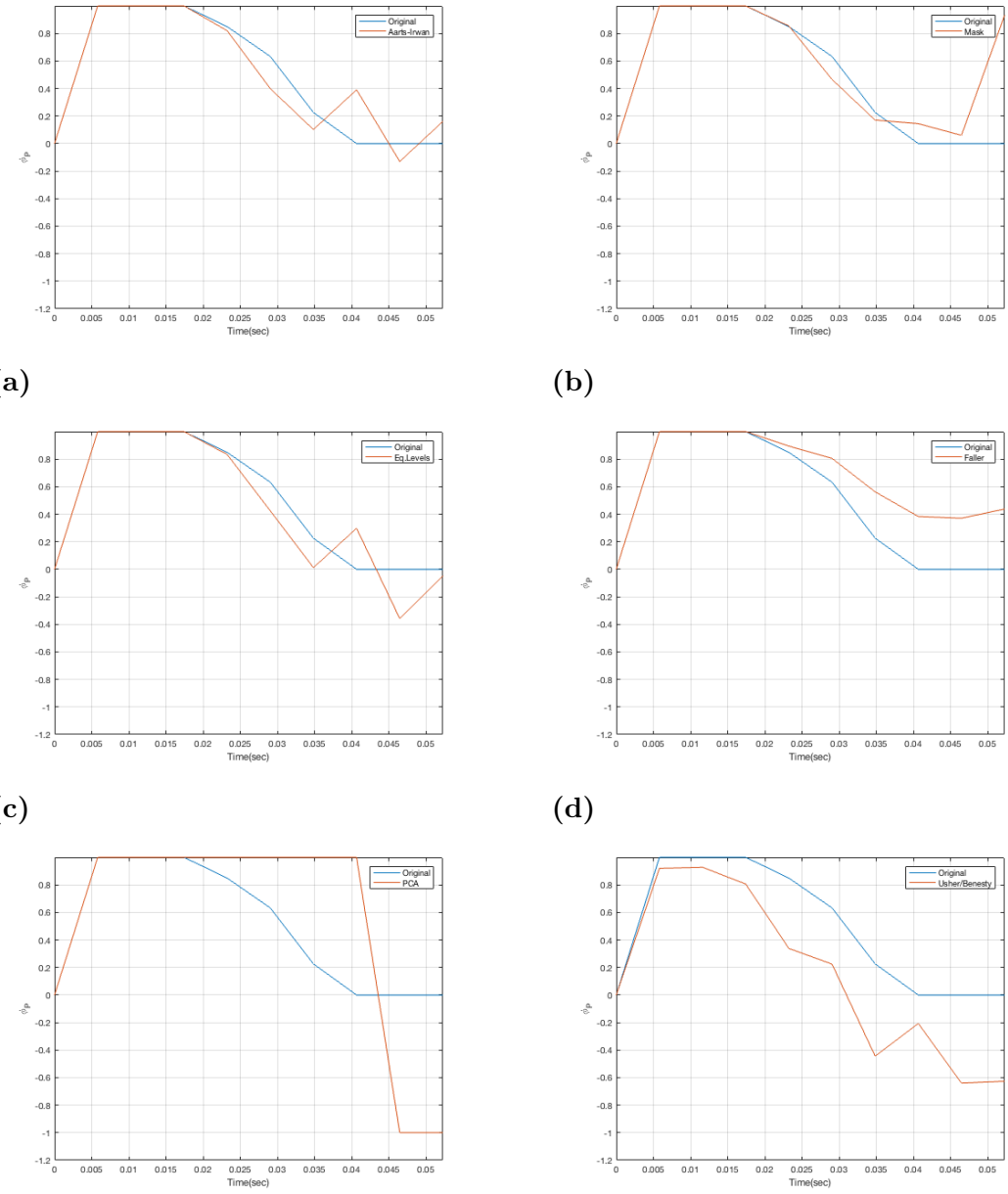


(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.25: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.4.4 Primary Cross-Correlation Coefficient, ϕ_P

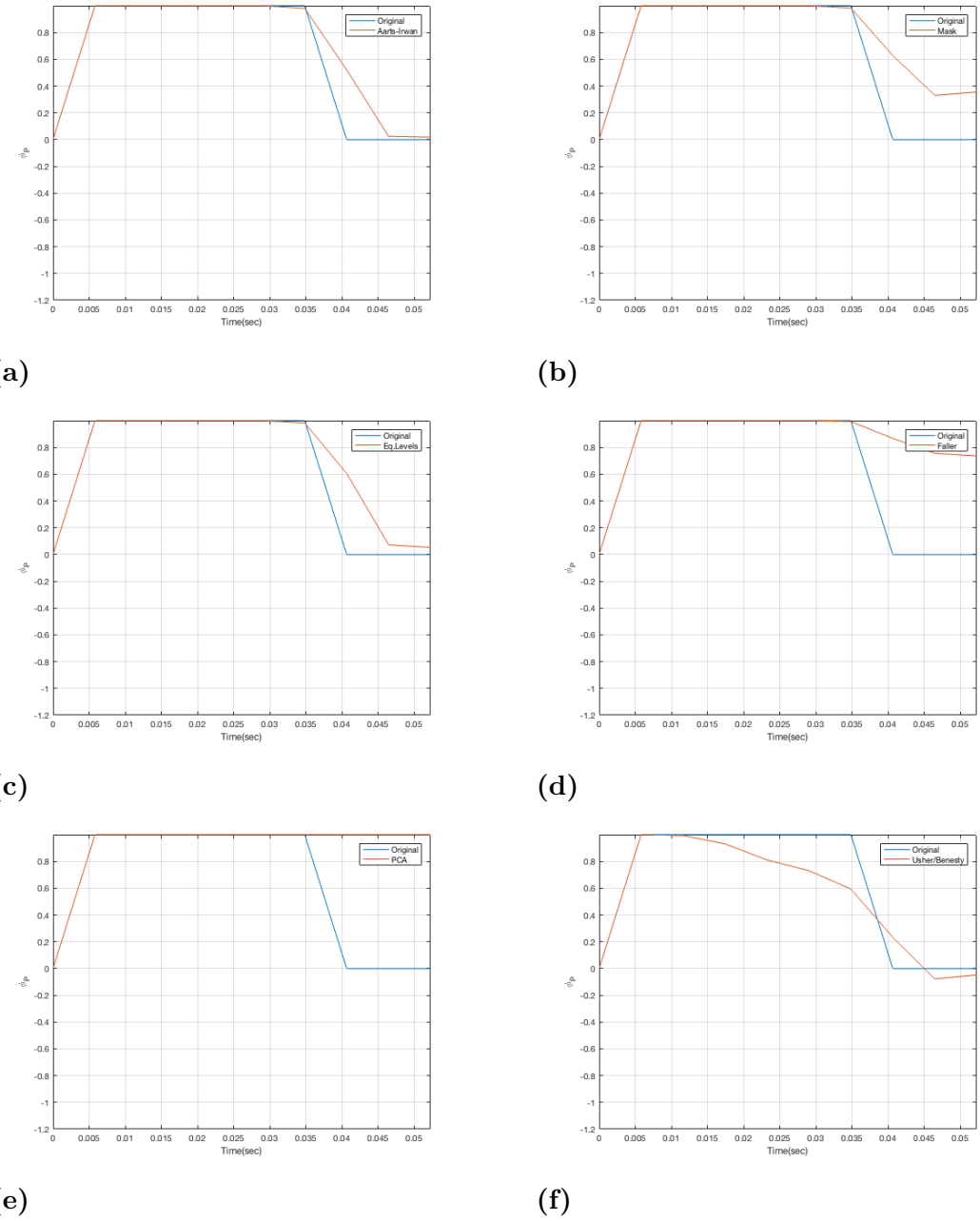
5.4.4.1 Transient Pulse



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.26: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.4.4.2 Gaussian Noise



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.27: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

Discussion

Figures 5.24-5.25 and 5.26-5.27 show the ambience and primary cross-correlation, given by equations 5.3 and 5.4 respectively. In the former graph, the mapping function and Equal Levels method produce ambience cross correlation close to zero, with values ranging from -0.1 to 0.1 for the Gaussian noise and higher fluctuation for the transient, yet uncorrelated signals. Usher-Benesty's method also produces

ambient signals with correlation close to zero (figure 5.24f), a fact which is more clear for the Gaussian noise (figure 5.25f) Faller's method and PCA generate ambience with negative cross-correlation. The explanation has already been given for the PCA case in section 5.6.3.4. For Faller's method, it seems that the high negative values of ϕ_A is a result of primary components with opposite phases leaked to the ambience. At the same time, the high ambience correlation produced by Aarts-Irwan method means that, the transient is not affected by the decorrelation filter applied, whereas in figure (figure 5.25a) the ambience of Gaussian noise is significantly decorrelated. Last but not least, in figures 5.26 and 5.25 the primary cross-correlation is depicted the first time instants, where the primary signal is initially present. The measure is correctly calculated for almost all the methods. Faller's and PCA also appear to estimate a part of the ambience as primary, in the Gaussian noise case (figure 5.25e). This is probably attributed to the insufficient energy difference between the primary and the ambience. At the same time the mapping function misinterprets a small amount of ambience of the transient signal as primary , reaching high levels of correlation at 0.06 s. However, this is probably residual primary signal with very low levels. On the contrary, as depicted here Usher-Benesty's method slightly underestimates the primary component, for both signal types.

5.5 Results of Commercial Recordings

5.5.1 Ambience Extraction

The results [1]ensued from the implementation of the up-mix algorithms are depicted in sections, from 5.1.1 to 5.1.6. The original left and right audio signals are plotted along with the elicited ambience for every method, in order to gain an overview of the magnitude of the decomposition. Although, the specific graphs are only available for an eyeball inspection, they directly reveal some of the systems' merits and weaknesses. In general, the derived ambience signals from all the methods have decreased amplitude compared to the original signal. That is actually what it is expected to happen, since the part of the signal responsible for the reverberance imagery have less amplitude than the direct components [44].

[1]Aarts-Irwan method extracts a monophonic ambient signal. In order to duplicate the channel, without maintaining exactly the same properties, decorrelation is applied in order to reduce the similarity of the signals. These techniques affect the phase of the signal, without altering its magnitude.

5.5.1.1 Mapping Function (Method of Avedano-Jot)

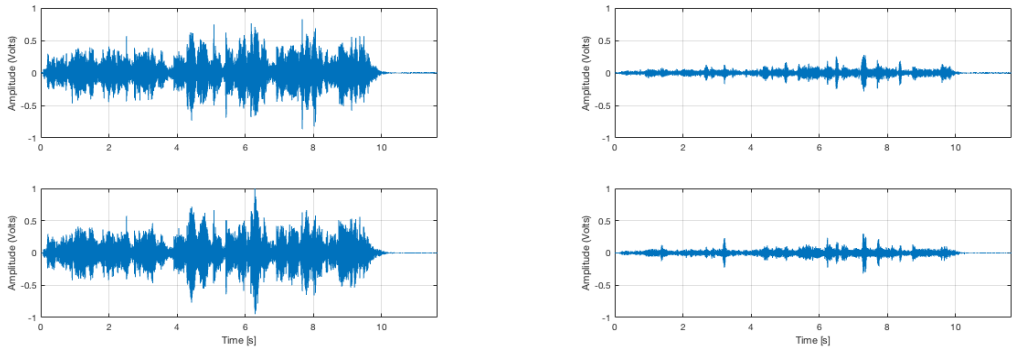


Figure 5.28: Initial congas left and right time signals and the corresponding extracted ambience.

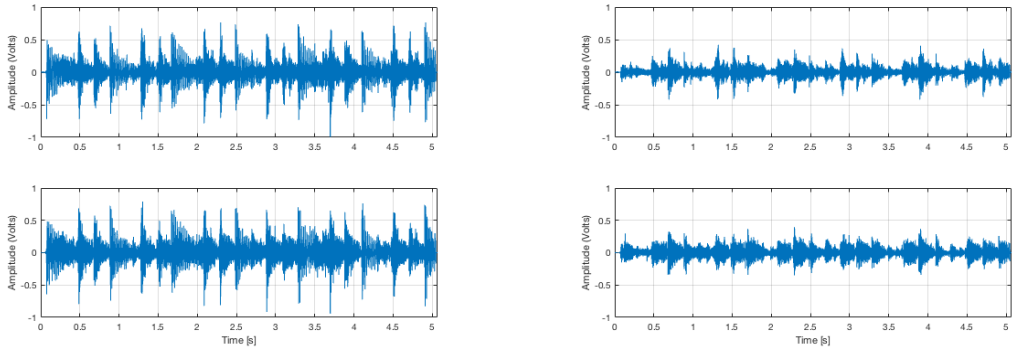


Figure 5.29: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

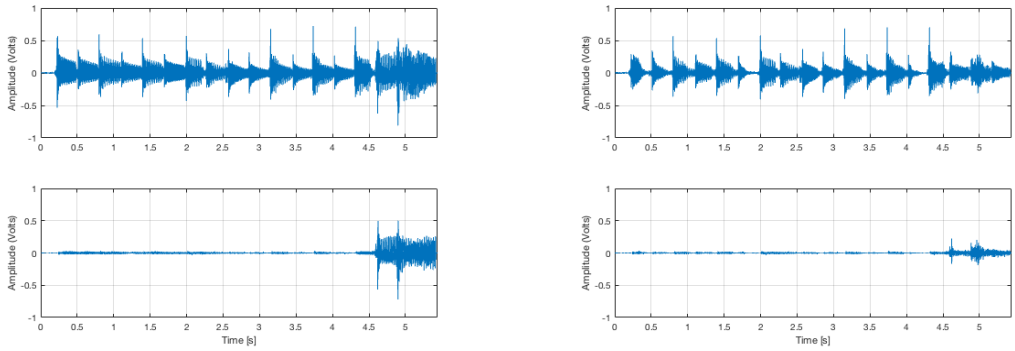


Figure 5.30: Initial electric guitar left and right time signals and the corresponding extracted ambience.

The most extensive problems are pointed out in the hard panned uncorrelated part of the electric guitar. The majority of the methods either give ambience channels

with very low amplitude (methods of PCA, Eq.Levels, Faller,Aarts-Irwan) in the channel containing the main source, or in contrast they yield signals having very high amplitude, as in Avedano-Jot and Usher-Benesty. In the former cases, all the methods produce at least audible results except PCA. The reason lies in the assumption of PCA, that the direct components are dominant in an audio mix. Hence, due to substantial more energy in the left channel, all the signal energy is attributed to the primary component. As for the methods overestimating the ambience, this simply happens because the direct components do not exist in the left channel and they cannot be canceled out, since the cross correlation assumptions are critical to the algorithms proper operation.

5.5.1.2 Principal Components Analysis

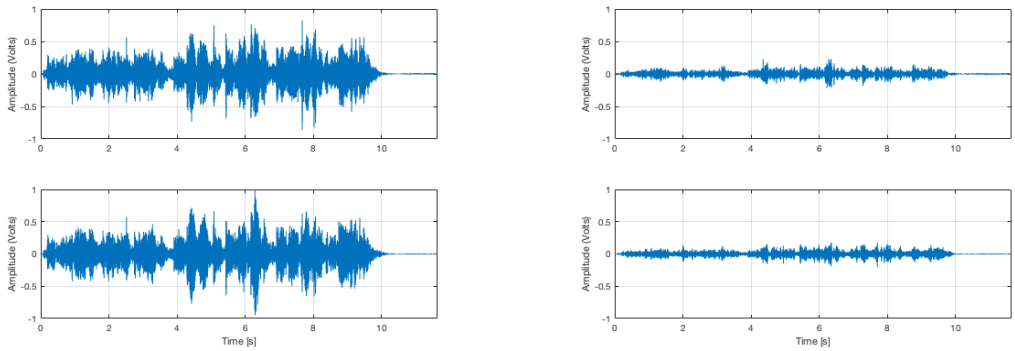


Figure 5.31: Initial congas left and right time signals and the corresponding extracted ambience.

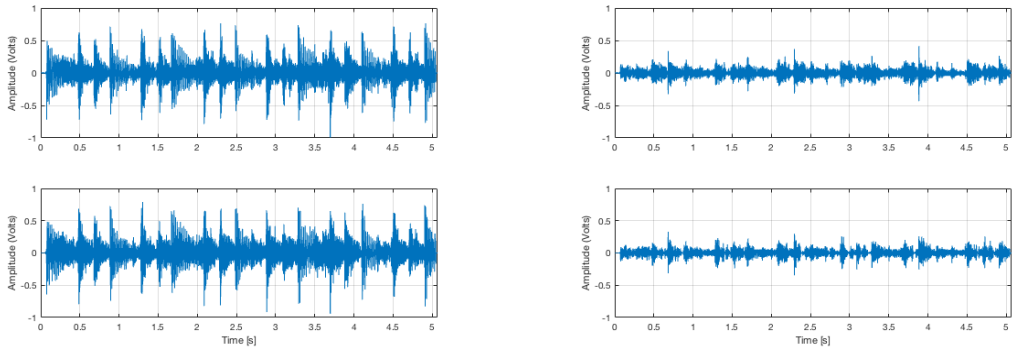


Figure 5.32: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

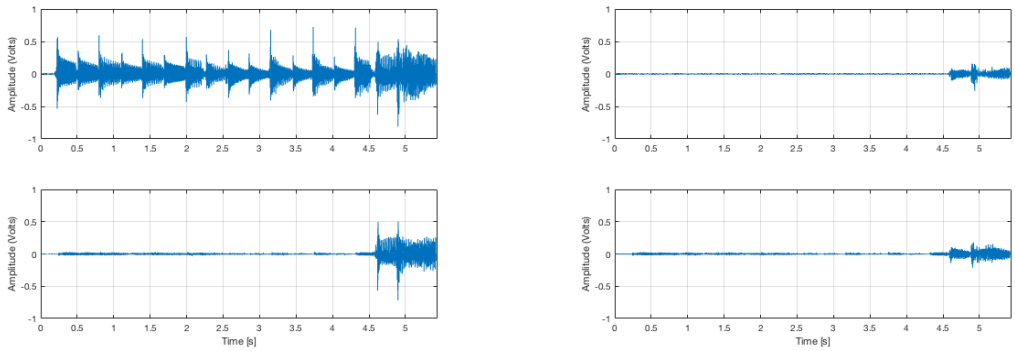


Figure 5.33: Initial electric guitar left and right time signals and the corresponding extracted ambience.

5.5.1.3 Scalar Ambience Extraction Masks (Equal Levels of Ambience)

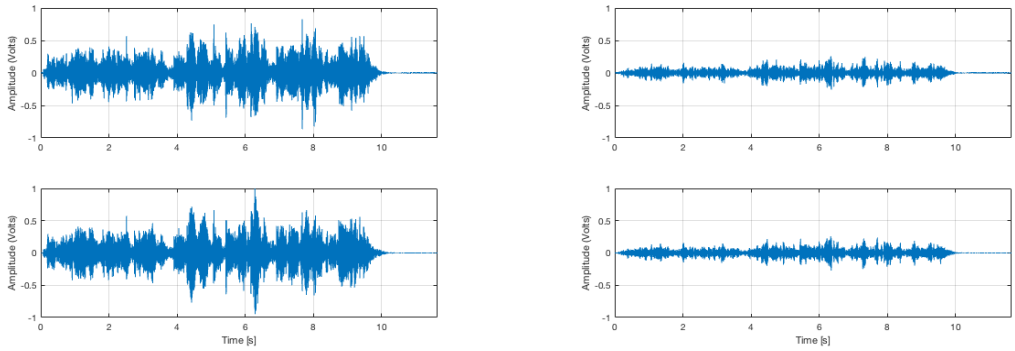


Figure 5.34: Initial congas left and right time signals and the corresponding extracted ambience.

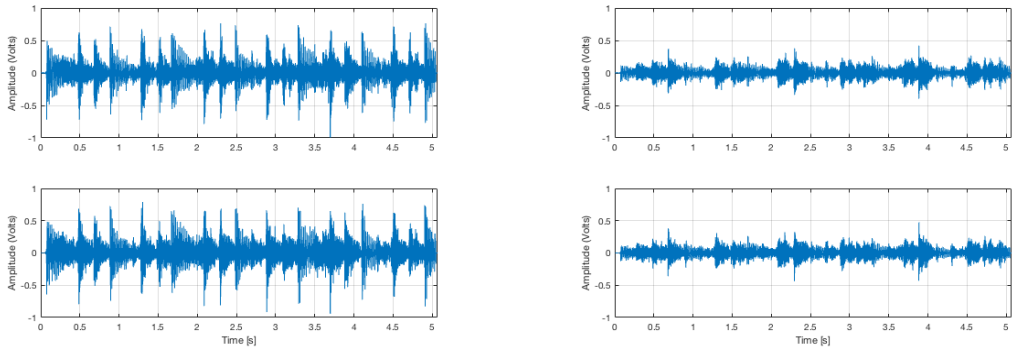


Figure 5.35: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

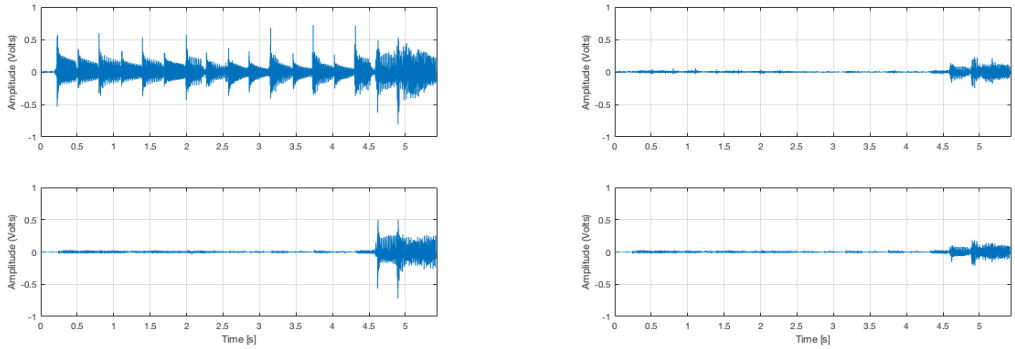


Figure 5.36: Initial electric guitar left and right time signals and the corresponding extracted ambience.

5.5.1.4 Least Squares Estimates (Method of Faller)

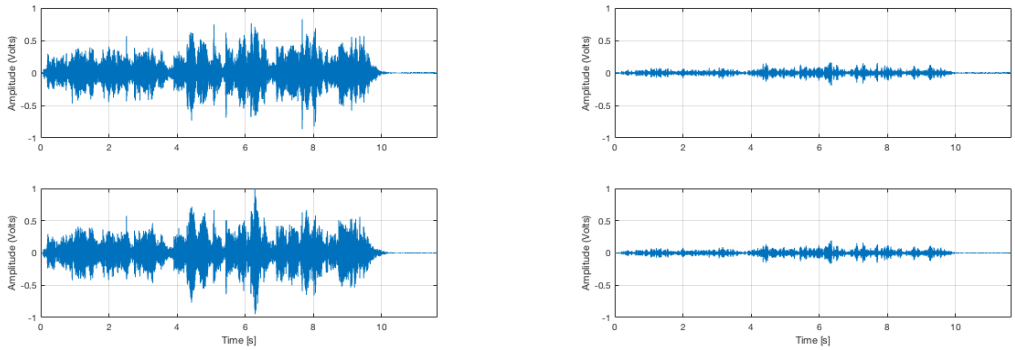


Figure 5.37: Initial congas left and right time signals and the corresponding extracted ambience.

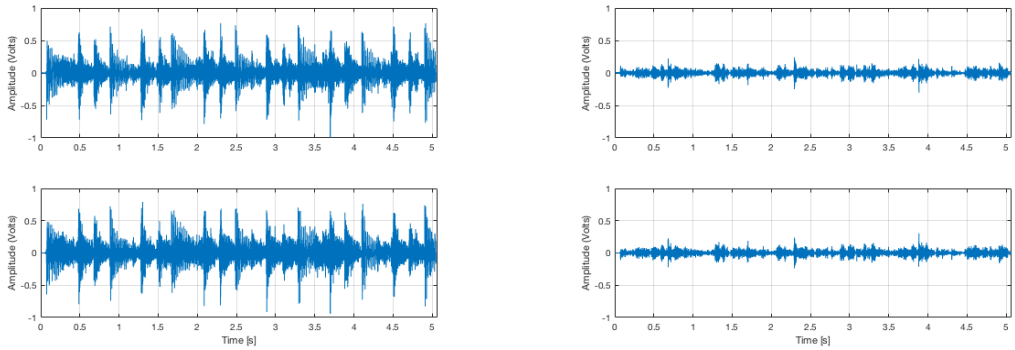


Figure 5.38: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

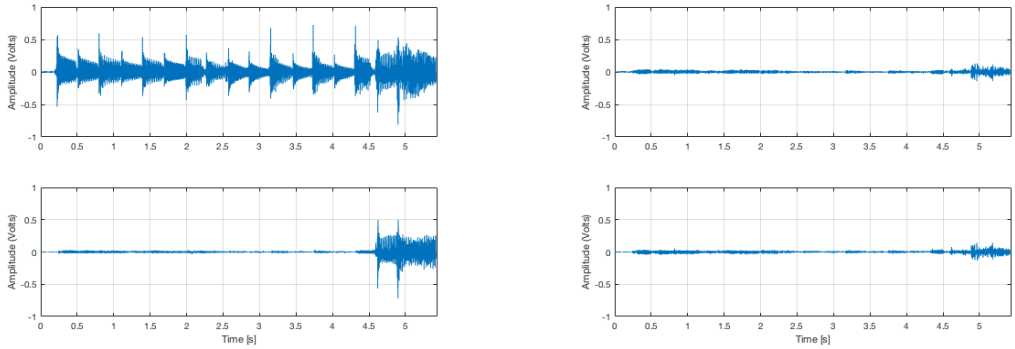


Figure 5.39: Initial electric guitar left and right time signals and the corresponding extracted ambience.

Apparently, all the upmix algorithms achieve low levels with a single source signal, but Aarts-Irwan's method seems to deviate from the general trend in figure 5.40, with the amplitude being even lower and very close to zero, separating only a certain amount of ambience. The reason is the high correlation of congas' channels (figure 5.1). It is attributed to the method's function, where fully correlated signals produce almost zero amplitude [20].

5.5.1.5 Time Domain LMS Filter (Method of Aarts-Irwan)

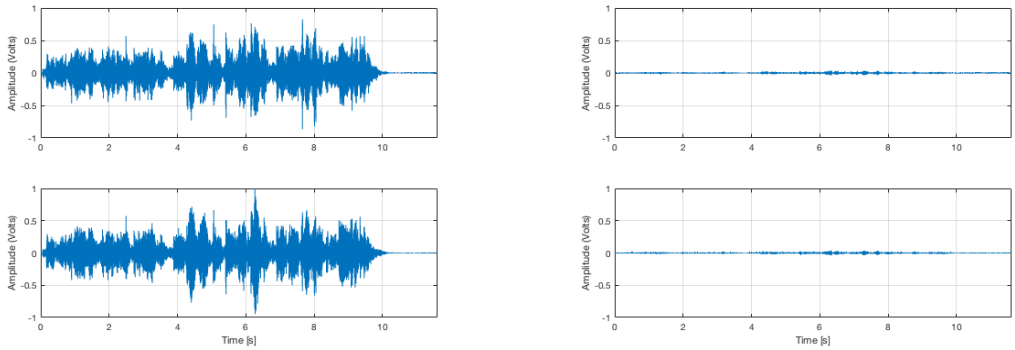


Figure 5.40: Initial congas left and right time signals and the corresponding extracted ambience.

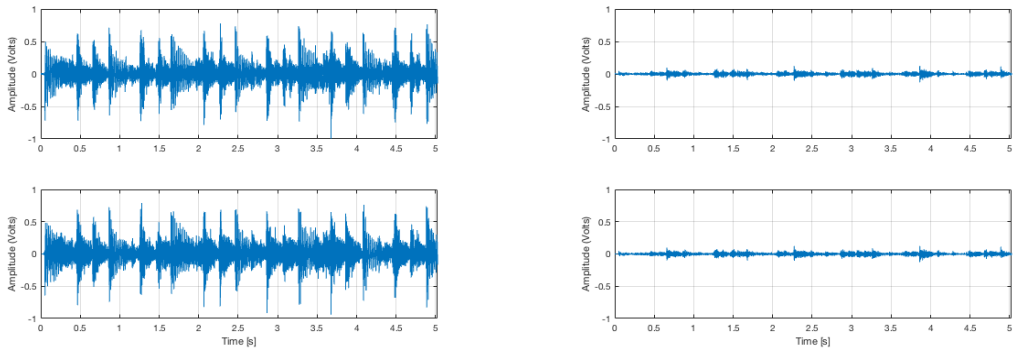


Figure 5.41: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

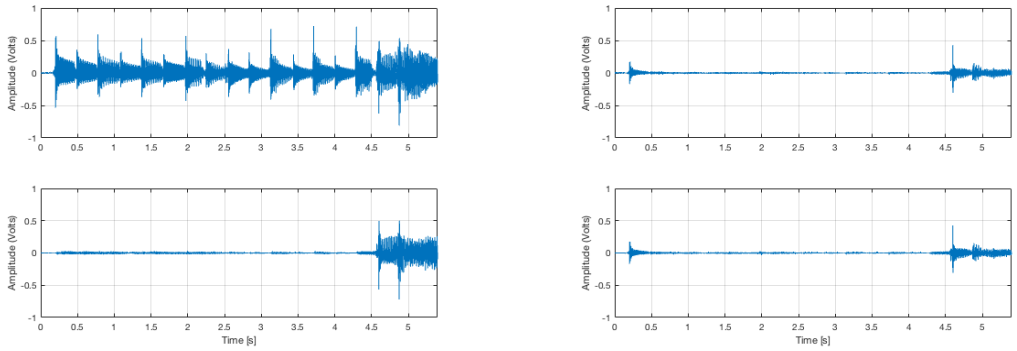


Figure 5.42: Initial electric guitar left and right time signals and the corresponding extracted ambience.

Regarding the ambience attributed to the pop-rock item, the amplitude is lower than that of the original music clip, but Usher-Benesty's method seem to yield a suboptimal separation. In figure 5.44 the ambience appears to have multiple prominent peaks, deviating from the relatively uniform visual impression, namely with little amplitude difference from the original channels, along several signal time instants. Therefore, the method doesn't lead to a sufficient separation, having as inputs signals with more than two sources [45]. On the other hand, with multiple sources and high correlation the method of Aarts-Irwan extracts reverberation of higher amplitude (figure 5.41), than with a single source. The latter possibly indicates that high correlation isn't necessarily a deterrent factor for ambience extraction. The system never extracts the ambience, when the stereo channels are fully identical. In cases with high correlation and utilizing the particular up-mix system, one derives at least a part of the ambience.

5.5.1.6 Normalized Least Mean Squares in Frequency Domain (Method of Usher-Benesty)

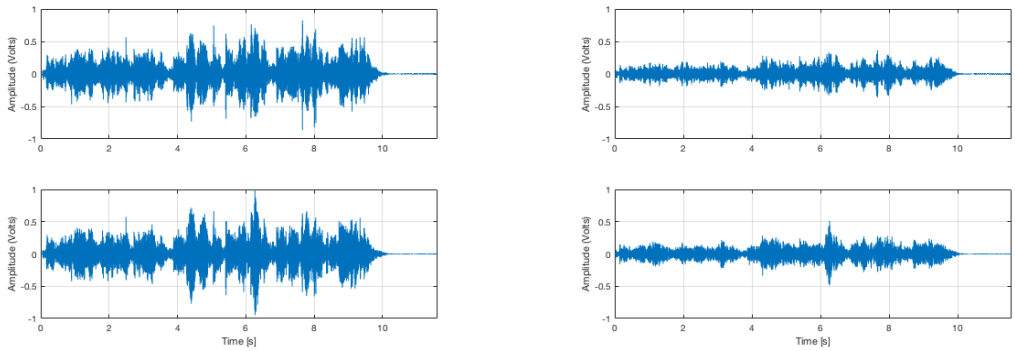


Figure 5.43: Initial congas left and right time signals and the corresponding extracted ambience.

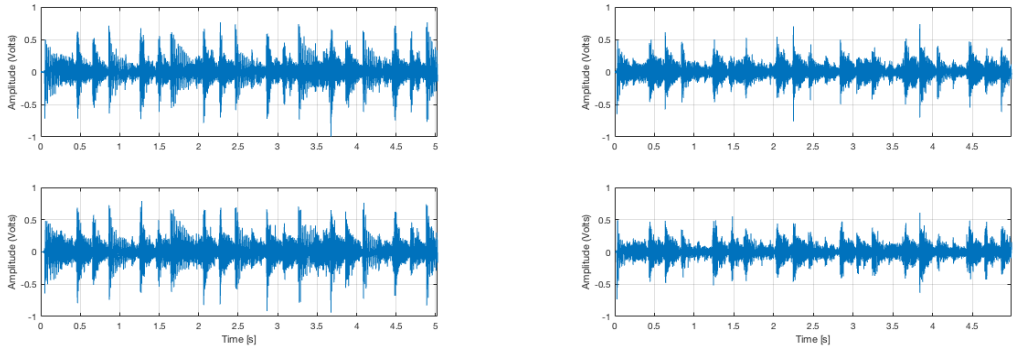


Figure 5.44: Initial pop-rock excerpt left and right time signals and the corresponding extracted ambience.

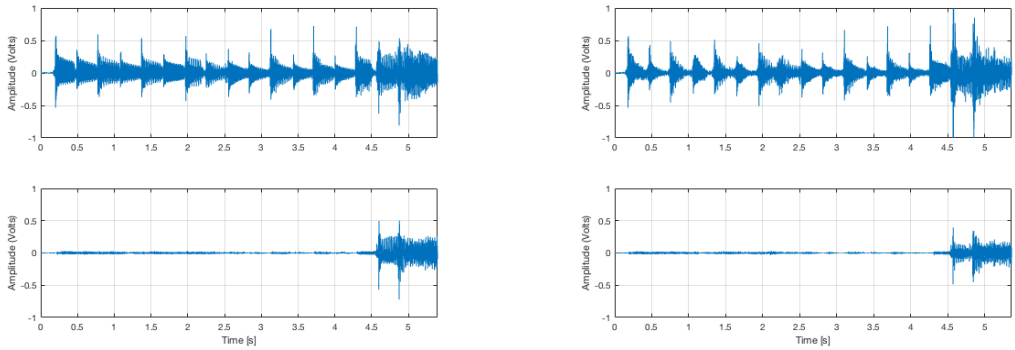


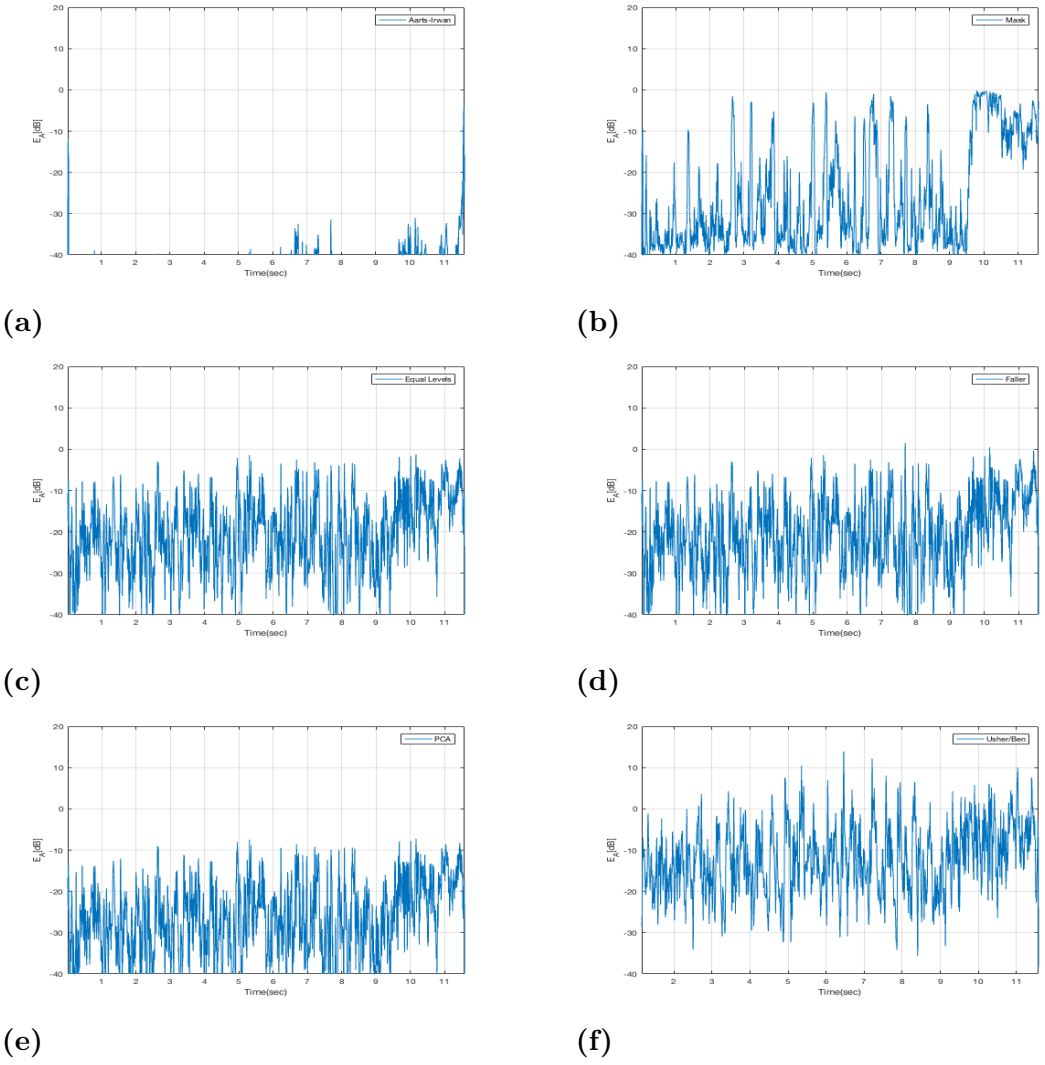
Figure 5.45: Initial electric guitar left and right time signals and the corresponding extracted ambience.

5.6 Evaluation Metrics For Commercial Recordings

The next subsections depict in graphs and analyze the algorithms' performance according to the fraction of extracted ambience energy ($E_A = \frac{||\vec{A_L}||^2+||\vec{A_R}||^2}{||\vec{X_L}||^2+||\vec{X_R}||^2}$), the cross correlation coefficients of the ambience ($\phi_A = \frac{\vec{A_L}^H \vec{A_R}}{||\vec{A_L}|| \; ||\vec{A_R}||}$) and the level difference ($L = \frac{||\vec{A_L}||}{||\vec{A_R}||}$), all defined in section 4.1. The results are obtained with the congas, the pop-rock excerpt and the electric guitar as inputs.

5.6.1 Ambience Energy Fraction, E_A

5.6.1.1 Congas-Single Source



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.46: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.1.2 Pop/Rock Excerpt-Multiple Sources

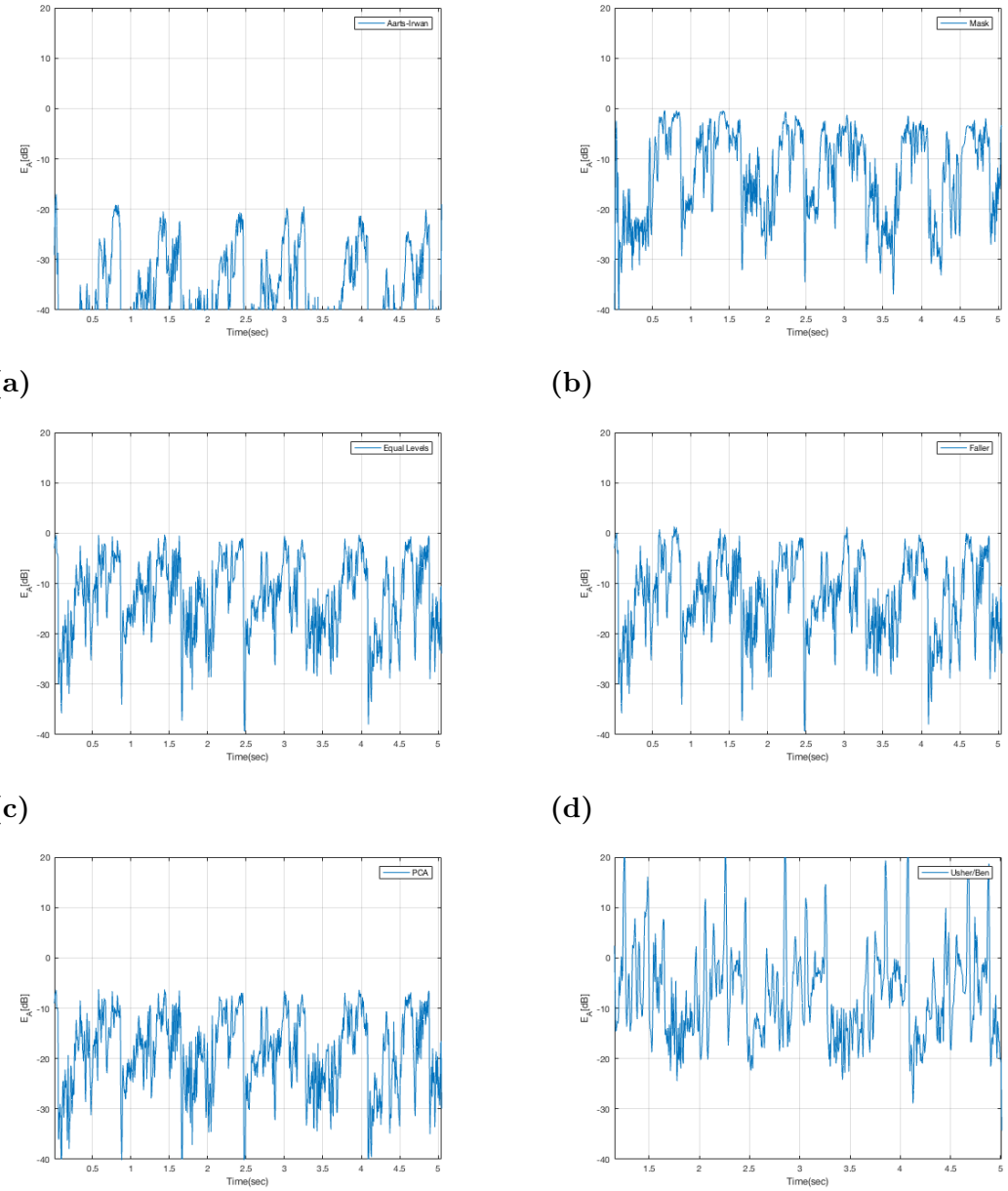
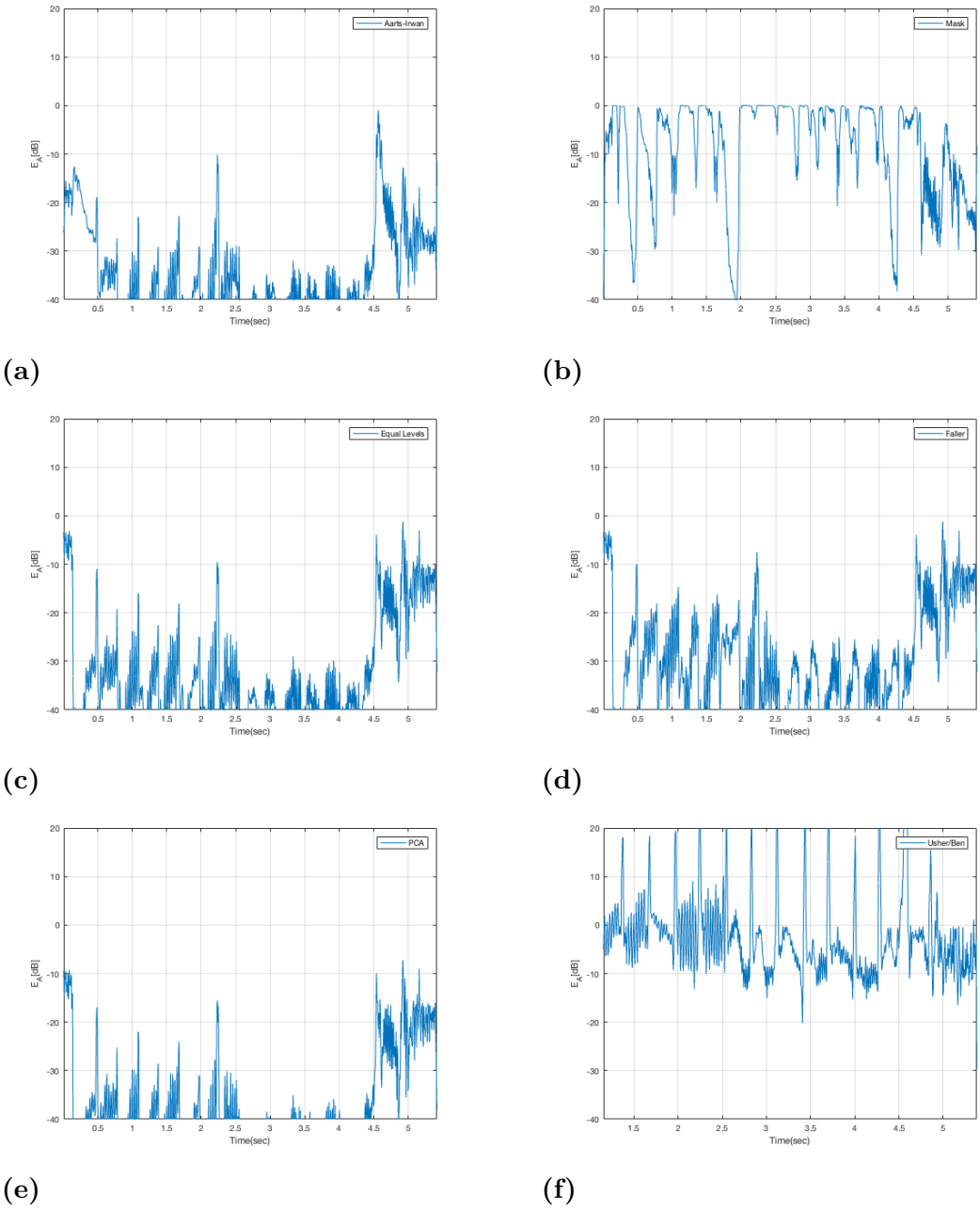


Figure 5.47: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.1.3 Electric Guitar-Uncorrelated Signals



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5.48: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

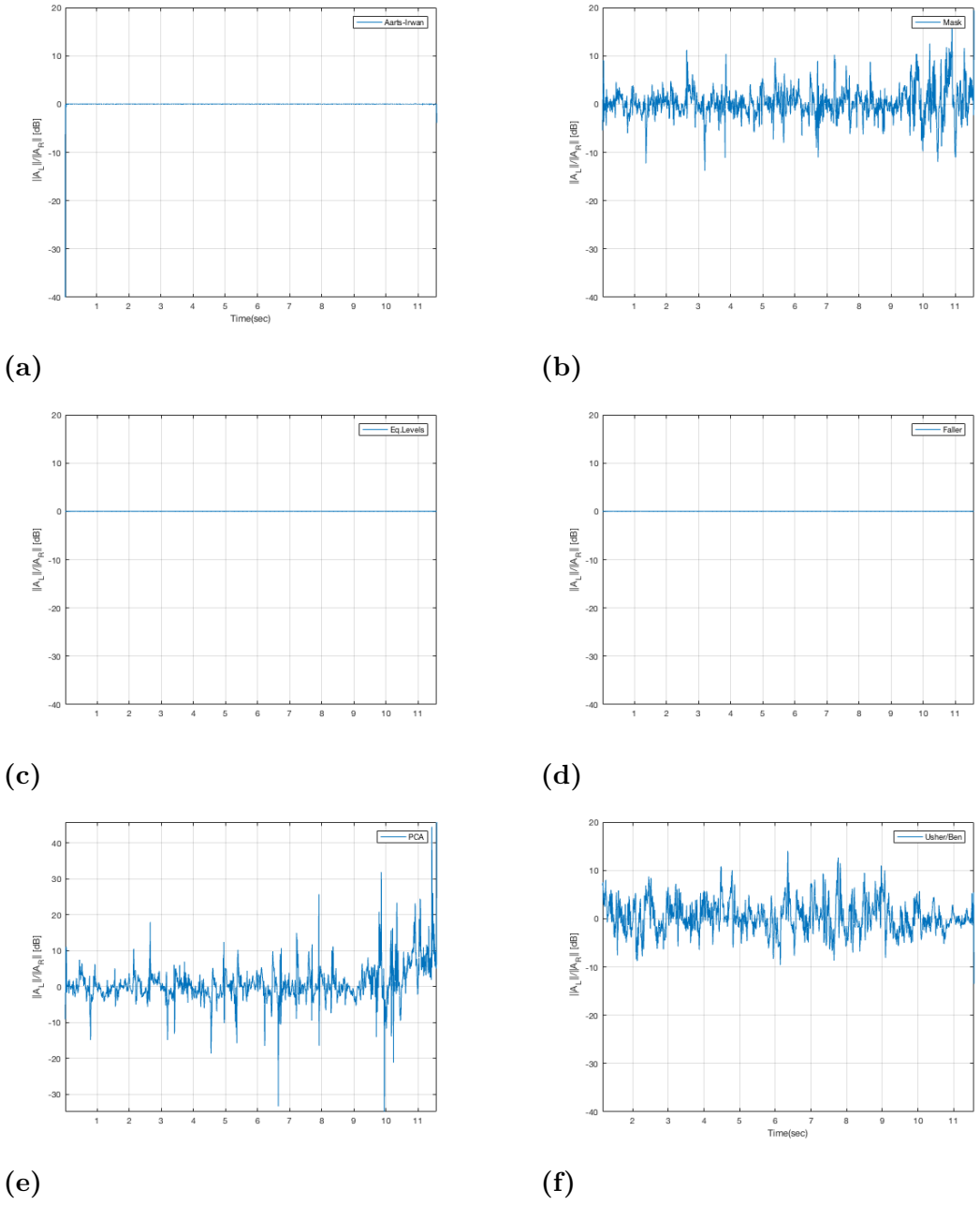
Discussion

In figure 5.46 the ambience energy fraction E_A curve fluctuates from -40 dB to 0 dB for most of the methods. Yet, the amount of ambience yielded by Aarts-Irwan is very low, i.e. below -30 dB, whereas Usher-Benesty slightly and partially over-estimates the ambience with a few peaks at 10 dB or slightly over 0 dB. It is also observed that PCA, Eq.Levels and Faller's curves follow approximately the similar

trend, with the last two having more or less the same magnitude. The mapping function, Eq.Levels and Faller graphs also have similar shape and levels in figures 5.47c,d and 5.48c,d. For the pop item, the algorithms seem to equivalently perform as in the congas. Moreover most of the curves follow a smoother trend than before, indicating probably a more clear distinction between the embedded primary and ambient components. However, even if Aarts-Irwan system extracts higher proportion of ambience than in the first signal case, the magnitude of E_A still is quite lower (10 to 20 dB) than the rest of the methods. PCA also never exceeds -5dB. On the contrary, Usher-Benesty overestimate the ambience in the signal with multiple sources, an artifact attributed, as it is reported, to the time-invariant convergence of the filter. For recordings with several sources active at different times, several adaptive filters could be utilized in parallel, as in echo cancellation [46]. For the last uncorrelated signal in figure 5.48, all methods extract a low level signal as ambience, but the mapping function and Usher-Benesty either extracting a large amount of the signal or wrongly computing the weights in the NMLS algorithm, respectively in each case. This happens, as the direct components don't exist at all in one of the channels. The latter will probably destroy the source image, transposing it more to the direction of the rear loudspeakers.

5.6.2 Level Difference, L

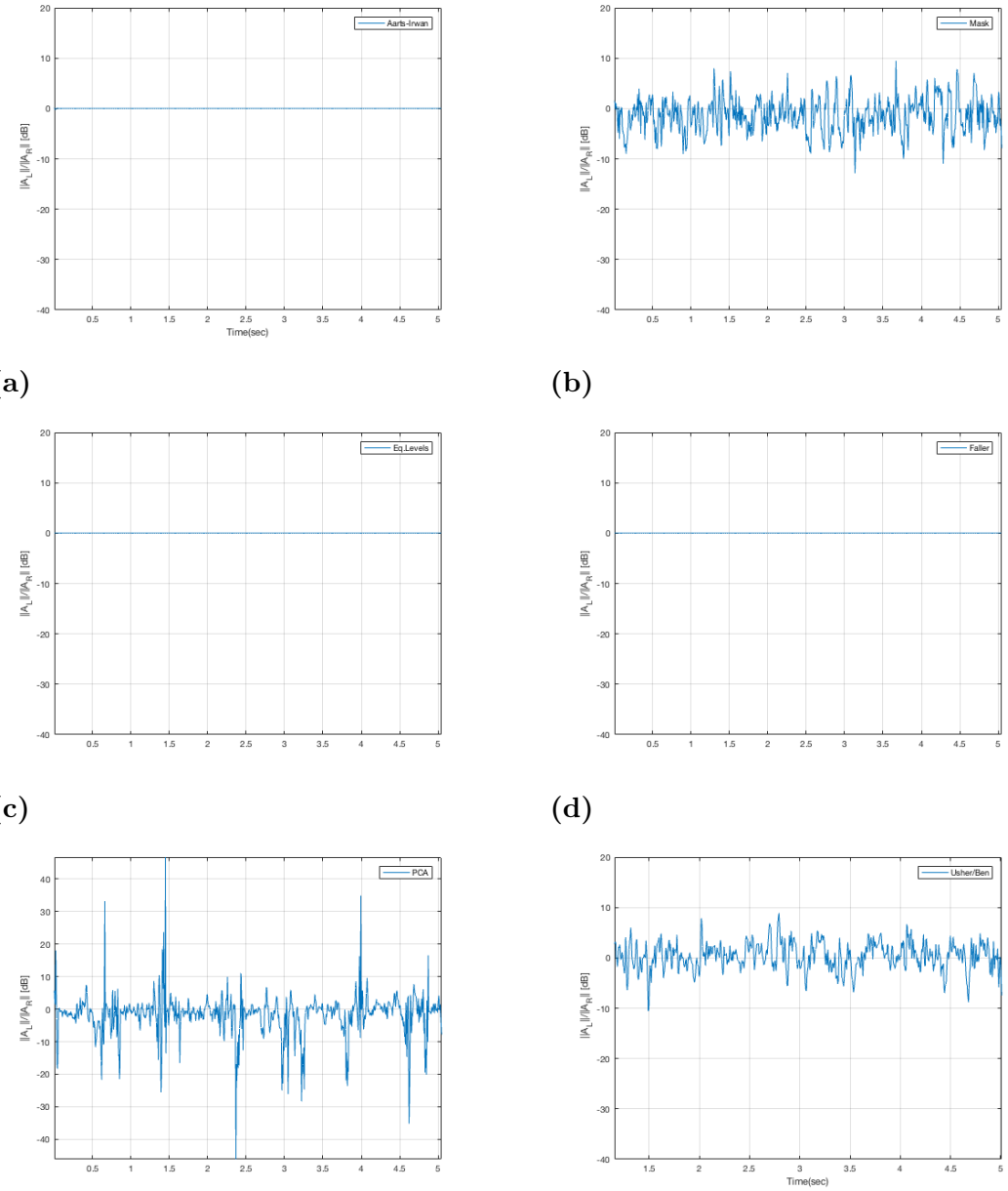
5.6.2.1 Congas-Single Source



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 5.49: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.2.2 Pop/Rock-Multiple Sources



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5.50: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.2.3 Electric Guitar-Uncorrelated Signal

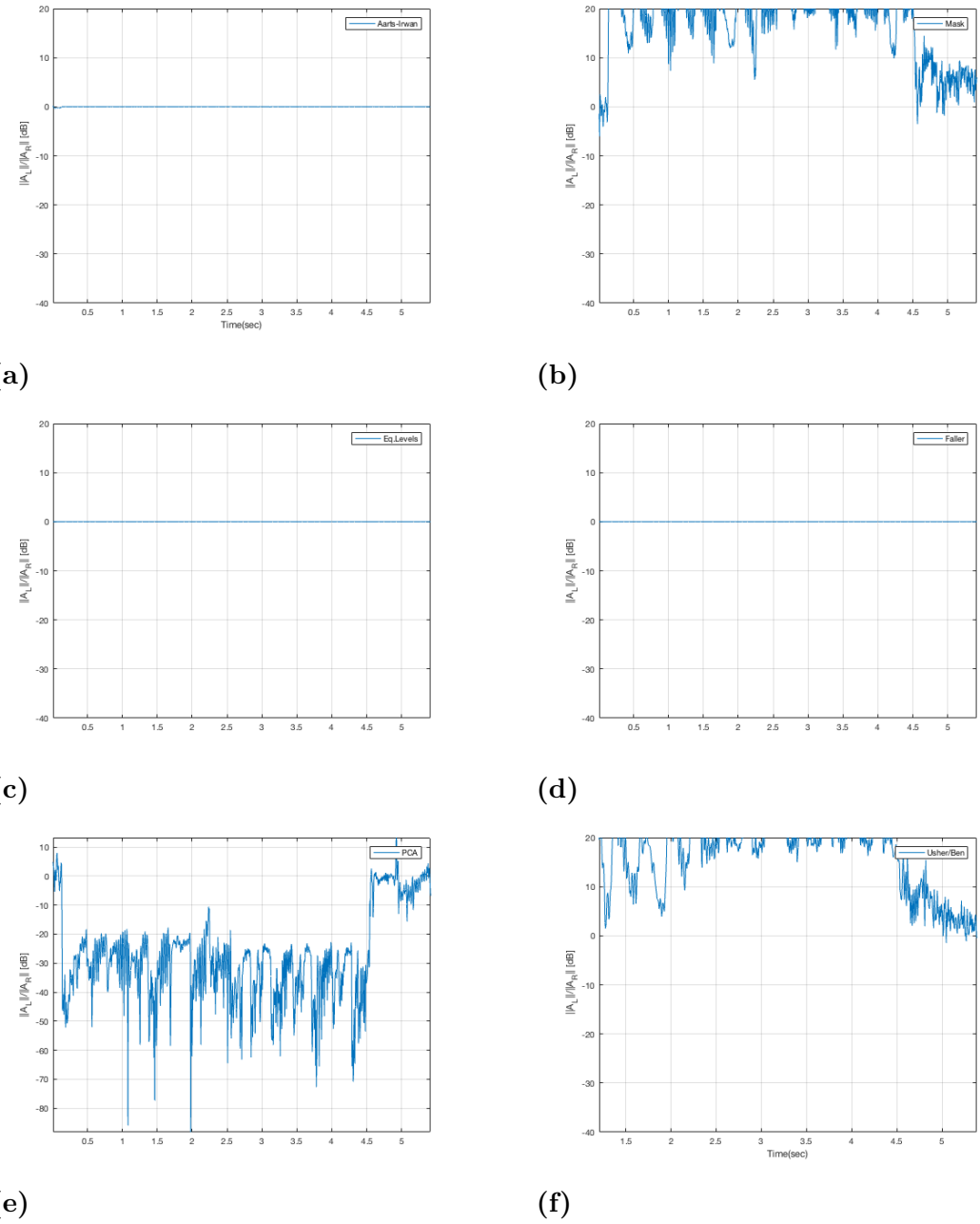


Figure 5.51: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

Discussion

Regarding the levels of ambience for all the types of signals, Aarts-Irwan, Eq.Levels and Faller's produce exactly the same amount of signal from every channel. In addition, the mapping function and NMLS (Usher-Benesty) graphs fluctuate around zero from -10 to 10 dB, meaning that in average the algorithms approximately separate equivalent amount of ambience in every channel (figures 5.49, 5.50). As

we have already described, the exception is of course the case of uncorrelated item, where they either remove the largest part or overestimate it, respectively. Thus the findings here is just a confirmation of what has already been inferred in previous discussion sections. At the same time,the PCA levels wiggling around zero (figures 5.49 and 5.50), but they occur very prominent peaks, implying that the ambience might also be left or right weighted. Hence, severe panning corrupts the stability of the ambience image [48]. Yet, this is to be expected, since PCA method is susceptible to mischaracterization of the ambience, if the signal doesn't fulfill very strict signal assumptions, initially defined in [22] and [19] and presented in section 4.1.

5.6.3 Ambience Cross-Correlation Coefficient, ϕ_A

5.6.3.1 Congas-Single Source

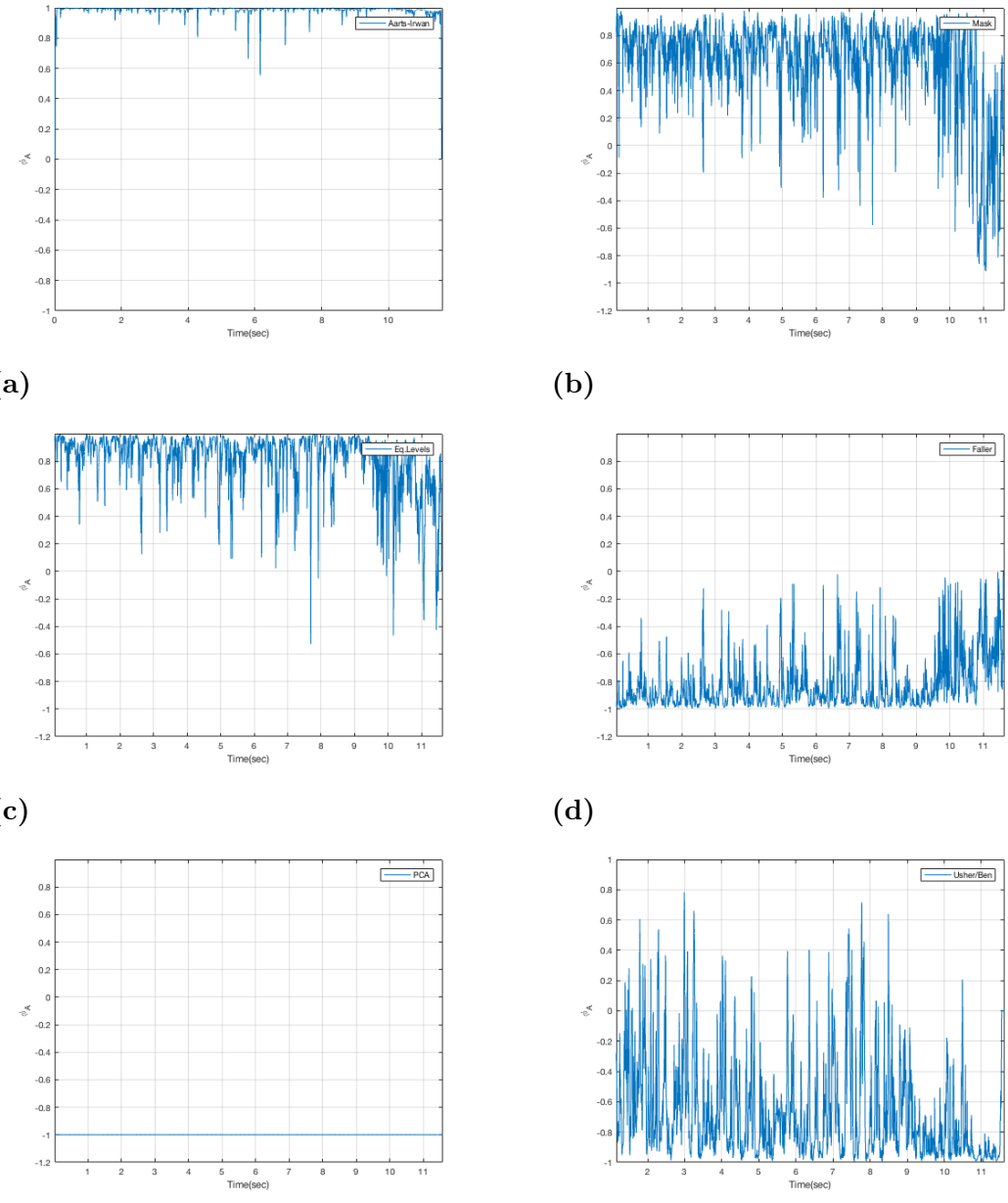
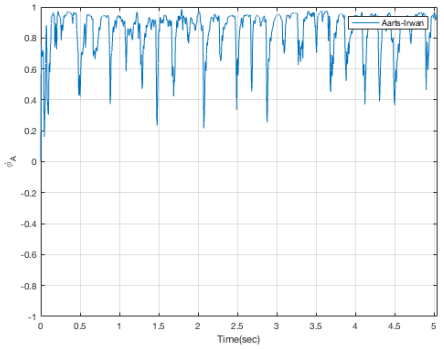
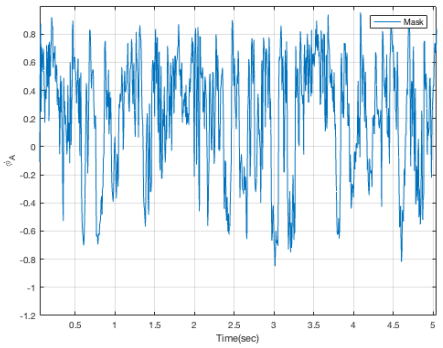


Figure 5.52: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

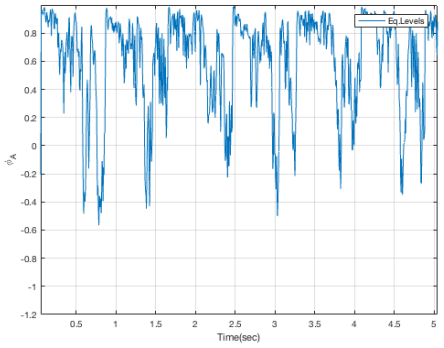
5.6.3.2 Pop/Rock-Multiple Sources



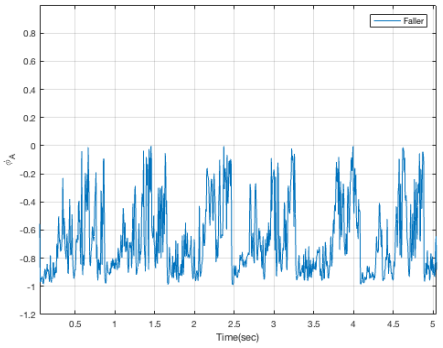
(a)



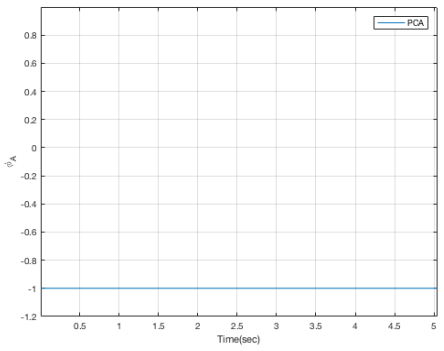
(b)



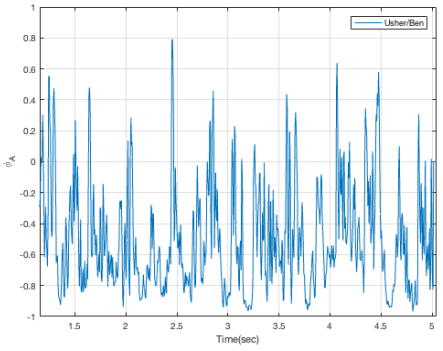
(c)



(d)



(e)



(f)

Figure 5.53: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.3.3 Electric Guitar-Uncorrelated Signals

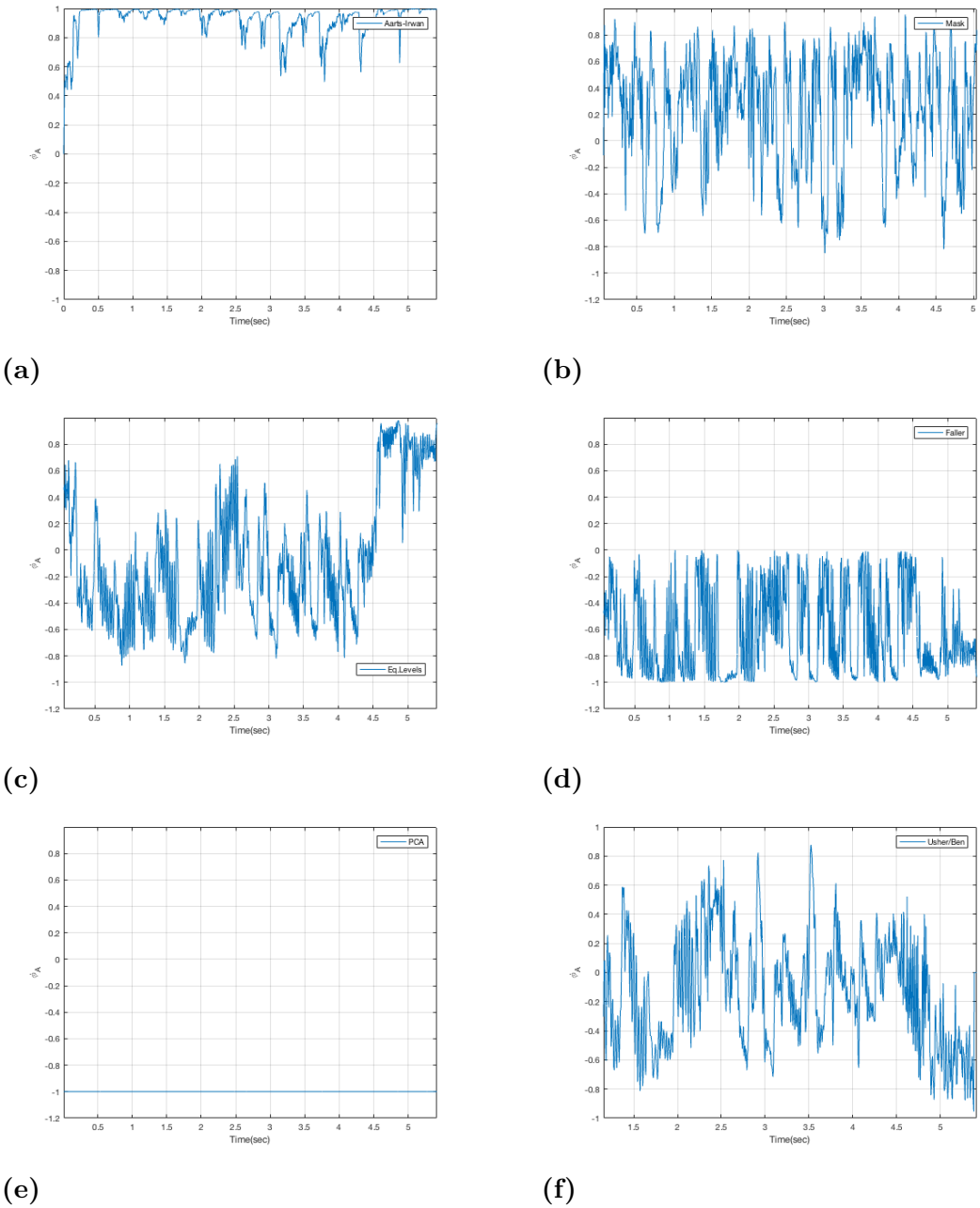


Figure 5.54: a) Method of Aarts/Irwan, b) Method of Avedano/Jot, c) Method of Eq. Levels, d) Method of Faller, e) Method of PCA, f) Method of Usher/Ben.

5.6.3.4 Discussion

Figures 5.52, 5.53 and 5.54 show the correlation of the channels containing the ambience, for every up-mix system. It is directly pointed out that PCA produces signals with negative correlation -1, because the signal space is formulated, so that the initial projections of the signal vectors on the ambience eigenvectors $\vec{u_1}$, (see the Method in section 4.4) lie in identical linear but opposite directions. Also,

Faller's method extracts ambience signals with negative correlation, a fact which is corroborated by the findings in [19].
On the other hand, the ambience cross-correlation of the method of Aarts-Irwan is high, with minor fluctuations close to 1 for all the input types. Yet, this is to be expected since the specific method yields only one channel of ambience and the second is filtered through a 4th order all-pass filter, as developed by Potard in [47]. A higher order all-pass filter could be used as well to achieve higher decorrelation, however over 8th order the amplitude outputs are significantly affected. As for the rest of the methods, not a clear image is available in respect to the ambience correlation. Thus, simpler input signal have to be used in order to draw a conclusion about them. As a result, a disadvantage of the objective measures emerges here. They are dependent on the capacity of the methods to extract the ambience and on the specific properties of the signals, which possibly pose several limitations for efficient ambience separation. Therefore, the correlation analysis, mostly investigates the up-mix systems susceptibility to deviations from the ideal signal forms.

6

Subjective Evaluation

A preference evaluation experiment is designed and undertaken, in order to validate the degree of appreciation of the discussed two-to-five upmix algorithms. The method employed and the final results are presented and discussed in the current chapter.

6.1 Participants and Stimuli Creation

This study looks into the effect of the six upmix methods on listeners preference and attempts to discuss possible reasons determining users judgment.

The experiment was carried out in the "Listening Room for Quality Test" at the department of Applied Acoustics at Chalmers University of Technology. It is considered a rather dry room with absorption treatment everywhere, except floor and reverberation time (RT_{60}) of approximately 0.5s. A conventional 3/2 loudspeaker configuration is used, according to ITU-R-775-3 standards and as shown in figures 6.1 and 6.2. The front and rear speakers are placed at $\pm\ 30^o$ and $\pm\ 120^o$ respectively, to the normal of the center. More specifically, the equipment used throughout the test is comprised by:

- Five Genelec 8030A studio monitors
- A Macbook Pro late 11 laptop
- Presonus Firestudio audio interface
- Cables

The loudspeakers are calibrated so as to radiate equal sound pressure levels at the listener's position, which is considered the sweet spot at 60 cm from every monitor. Since a part of the room is used for the experiment, the conventional value for speakers emitted levels(78±2 dB) in [55] is considered too loud for the allocated space. After an extensive listening, the SPL value is adjusted and attenuated to 65 dB. This seems to create a more natural, homogeneous and smoothed acoustic experience for the users, who were sitting on a non-rotating chair throughout the listening test.

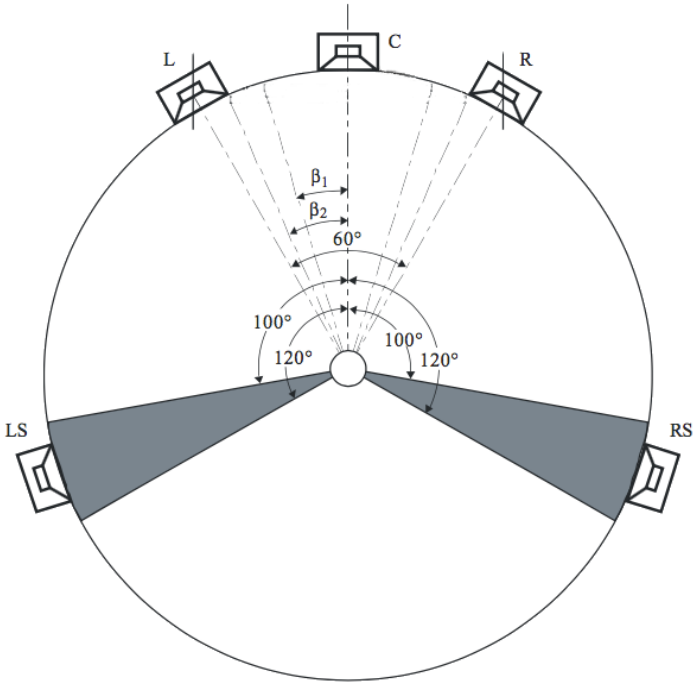


Figure 6.1: Arrangement with three frontal and two rear loudspeakers, taken from [1]

Figure 6.2: 3/2 setup during the experiment.

Three different stimuli are used as input sources for the upmix algorithms. An excerpt of trumpet, congas and male voice speaking. Although, they are not representative of a wide range of potential stereo recording items, their simplicity is believed that sufficiently unveils the advantages and downsides of each algorithm. The stimuli are recorded in the reverberant room (figure 6.3) at the facilities of Volvo Cars in Torslanda in Gothenburg (Sweden), by playing back the material

through speakers and then recording them using X/Y microphone technique. The material used is anechoic recordings or at least dry recordings, hence any additional reverberation is only added by the microphone setup and the reverberant room.

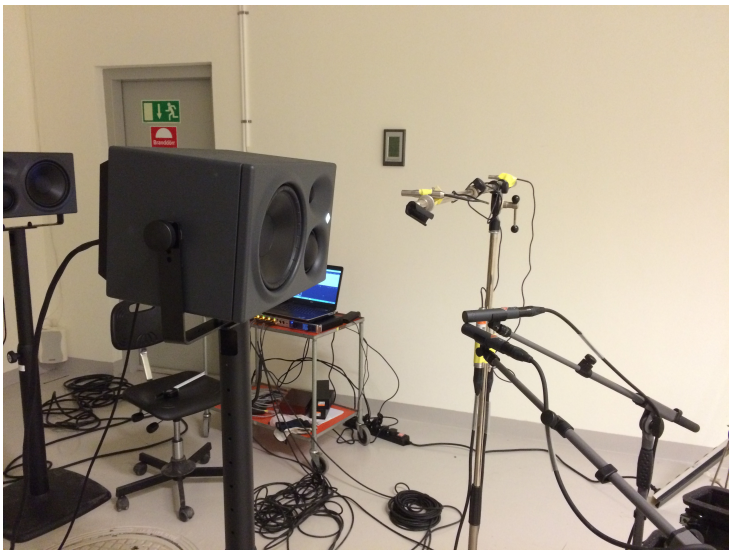


Figure 6.3: Photo taken from the recording session.

First, the items are recorded with two speakers at 1.5 m each to the microphones and 1 m distance between each other. Then, the same thing is recorded again with one speaker, which is moved closer to 30 cm from the microphones to get more direct sound in the mix.
Nineteen people took part in the experiment. Some of them are considered experts, who are able to describe an auditory event in detail and they have a background in implementing and participating in subjective evaluation tests. However, the vast majority are either first or second year postgraduate students at the Division of Applied Acoustics, at Chalmers University of Technology, who had already experienced at least once a listening test session. All of them have reported no hearing impairments.

6.2 Method

In order to reduce the complexity and the errors of manually collecting the answers from the participants, and to better control the test, a software interface is created in SuperCollider 6.4.
The subjects could switch between stimuli and each time the excerpt is paused and not stopped, providing the opportunity to carefully listen and judge. Once the stimuli have been listened, the participants could express the magnitude of their preference by setting the handle of a horizontal slider accordingly. Moreover, the value of preference of every judgment is displayed in a small rectangle next to the slider. In case of indecision, a "No preference" choice is allowed as well, positioned in the middle of the bar. The slider represents a continuous scale of numerical values,

which spans from -5 to 5 and it allows for more elaborate analysis than a forced binary stimulus selection. [56].

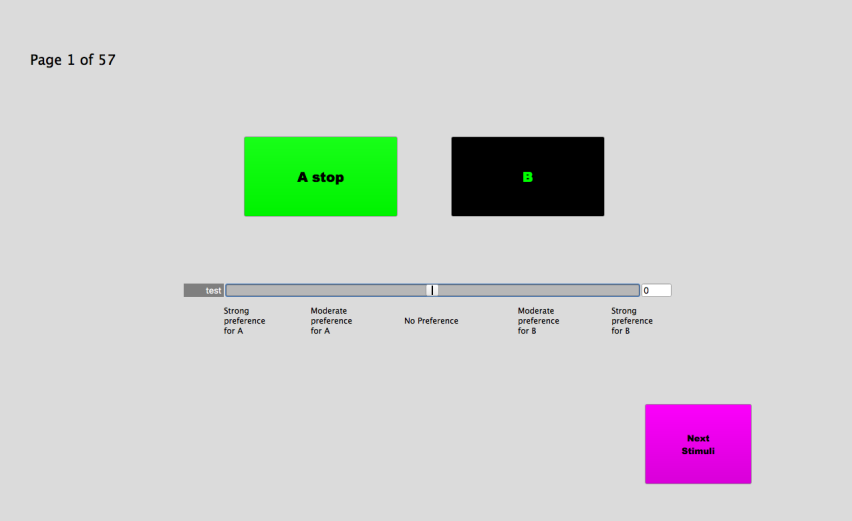


Figure 6.4: Enabled interface of preference evaluation

The audio stimuli were presented to listeners according to paired comparison framework. It is favoured over other, mainly for its simplicity as it facilitates the procedure for participants and at the same time allows for direct comparisons. In general, reducing or avoiding participants fatigue was a a rule of thumb kept while designing the test. On the other hand, this very rule also sets several limitations related to the duration. For instance, using more sources as stimuli could be more time-consuming and cumbersome, whereas reducing them would possibly yield, at least, less reliable results. Thus, each stimulus duration is kept 15 s long, playing in a loop, unless the participant decides to stop it or proceed to the next pair .
The experiment consists of three parts determined by the type of stimuli each time. The three audio items were processed by each of the six described up-mix systems and for every stimulus fifteen unique pairs are formed, namely $3 * 15 = 45$ unique pairs, hence 45 collected answers in total. A training session of four pairs is preceded every stage of the experiment, so that the participants could get a broad idea and familiarize themselves with possible extreme and very similar cases, reflecting the full preference scale of the slider (see in Appendix A.2). As a consequence, the total number of pairs, determining the duration of the experiment is $4 * 3 + 3 * 15 = 57$ pairs per participant. In average, the time taken to every listener to fulfill the test lasts from half an hour to one hour.

6.3 Results

6.3.1 Preliminary investigation

In figure 6.5, box plots and the spread of participants' responses are displayed, in order to acquire a relatively clear overview of the scope of the answers, denoting preference for every up-mix method. The box plots show the median values, the

interquartile range (upper and lower quartile) and possible outliers (e.g. in "Usher/Benesty" box plot). The median values are mostly negative on the scale range (from -5 to 5). However, this is attributed to the position each stimulus had in every pair, because they are randomly assigned to buttons "A" and "B".

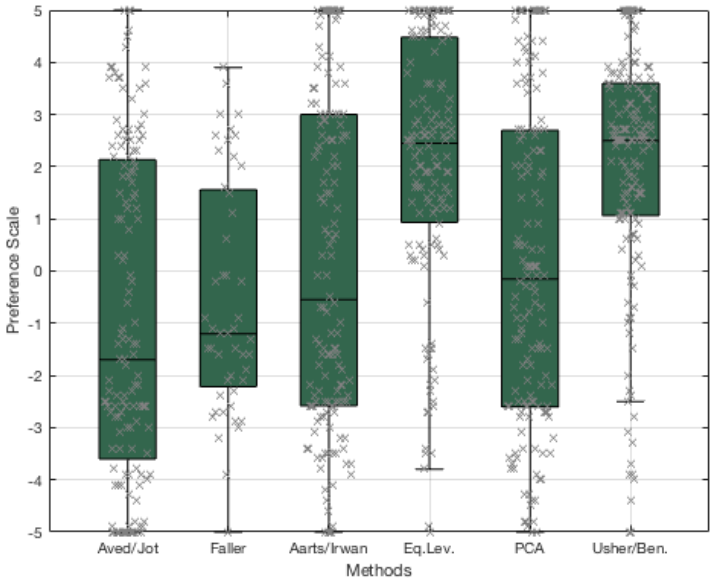


Figure 6.5: Box plots of accumulated responses, indicating preference for every up-mix method. Every answer is the outcome of the pairwise comparisons of the systems investigated here.

The responses in figure 6.5 represent the preference for every method over another, resulting from the comparison judgments of the pairs. It can be observed, that the preference answers in most methods cover the whole scale range except for Faller's. A slightly more confined box plot occurs, with significantly reduced density and number of answers. Another observation one could make is that there are answers of preference very close to zero in every box plot, especially in PCA. In that case, all the answers between -0.5 and 0.5 are discarded, because either the "No preference" choice is ill-defined, due to lack of a certain mark displaying the middle of the bar on the interface, or they are not considered a clear appreciation for a certain method. Nonetheless, these indefinite data doesn't seem to corrupt the results, as they are very few. The presented ones will be used to elicit the final preference scale.
In addition, the data are decomposed to answers per method and session, plotted in figure 6.6. The latter shows the level of potential dependency of audio items upon the up-mix methods, namely which way the spatial version quality of each stereo recording is affected by processing it with every algorithm. Faller's box plots display high inconsistencies, with the trumpet and voice session gathering none and very few moderate preference choices respectively, whereas significantly more participants choose this system during the congas session, with answers covering almost every preference scale. Moreover, discrepancies appear at Eq.Level's method, with the congas responses being limited to a part of the preference scale and covering mostly

the moderate area of the scale. Regarding the rest of the methods, the responses appear to have a higher degree of homogeneity among the different parts of the experiment, with box plots following approximately the similar trend.

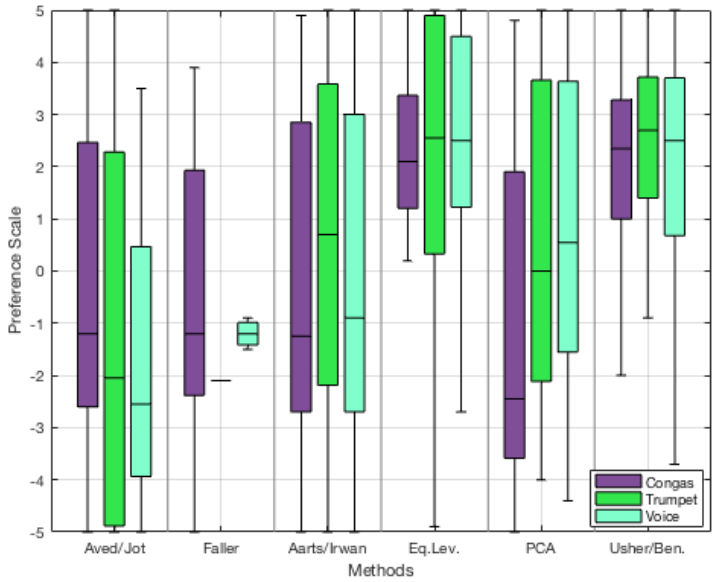


Figure 6.6: Box plots of accumulated participant responses, indicating preference for every method and stage of the listening test.

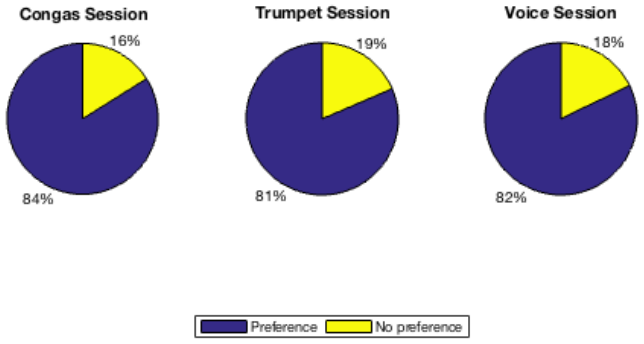


Figure 6.7: Percentage of the "No preference" responses, over all participants, for every session of the listening test.

To complete the preliminary investigation, the pie charts in figure 6.7 depict the percentage of "No preference" judgments throughout every stage of the experiment,

pool over all participants. Thus, it is apparent that the percentage of "No preference" choices is approximately the same either looking at congas, trumpet or voice part of the test. They also are just below 20%, which doesn't seem to corrupt the results with many "No preference" selections. This could likely be an indication about very identical listening experience or an implication about severe artifacts induced by some methods, posing difficulties to participants to make a choice. It must be underlined that only the zero ratings are kept, because they directly denote clear indecision. The values very close to 0, such as -0.1 or 0.1 and -0.2 or 0.2 are discarded for accuracy.

6.3.2 Mean Absolute Difference Between Judgments

In every stage of the experiment there are fifteen unique pairs of up-mix methods and each one of them is repeated three times for every participant. Thus, in figure 6.8 the average absolute difference of judgments is displayed, along with the respective 95% confidence intervals, so as to identify potential deviations of participants' scores and in which extent the methods' impact on listeners' preference remain unbiased by the stimuli. For instance, in case values are very small, then there are insignificant deviations between judgments and each method scores on average similar scores. The horizontal line represents the mean over all participants.

It is remarkable, that only the 19^{th} participant has a value slightly higher than the subjects' mean. Moreover, a relatively stable and global rate of changing preference is formalized, since the confidence intervals are small enough and approximately of the same width, with the exception of those belonging to participants P7, P9, P12,P14. Nonetheless, there is a moderate rate of changing selections, which underscores that the preference varies for different stimuli up-mixing with the same method. Still, the results maintain a certain degree of independence from the used stimuli, since the average is in the middle of the scale. As the percentage of no preference responses is relatively low (below 20%), it is pointed out that the outcome is mainly based on preference responses.

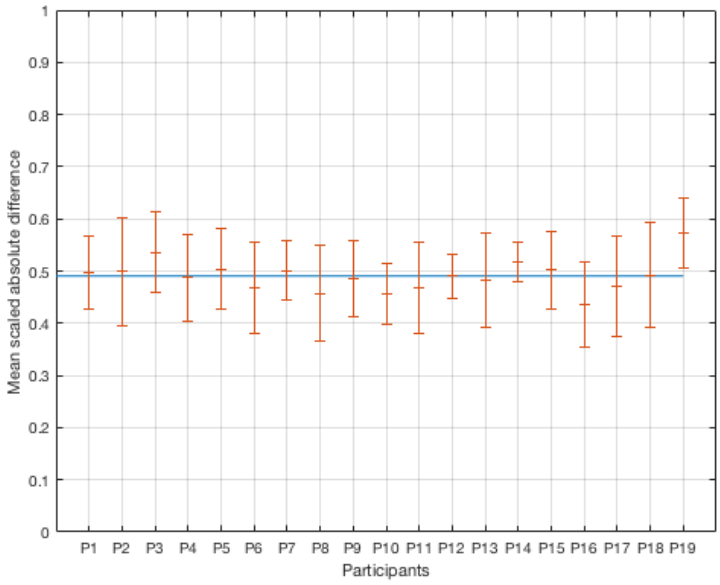


Figure 6.8: Average absolute difference and the corresponding 95 % confidence intervals of judgments, over all method pairs and sessions, for every participant.

6.3.3 Preference Scale for Up-mix Methods

The law of comparative judgments [57] and especially maximum likelihood method [58], assuming Thurstone's Case V (see Appendix A.3), is used to extract the scaling of preference. The data are processed according to Matlab code introduced in [59]. In a nutshell, the law of comparative judgments assumes that every potential choice is a Gaussian random variable and Case V additionally assumes all options have identical variance σ_i. Then, the magnitude of preference scale is derived by the number of times a stimulus is chosen over another.

The corresponding 95% confidence intervals for each of the preference scale values are calculated in a similar way as in [60], which follows the bootstrapping technique [61]; the scale is calculated several times for every sample of data selected from the original set. Eighty iterations are performed and each one of them consists of slightly less than 50% of the total data.

To be more precise, in a session being represented by a type of stimulus, e.g. congas, the number of judgments for a pair of upmix methods, say A-B, are as many as the entire number of participants, i.e. 19. From these total amount of data, 9 answers are randomly selected (that's why less than 50% of data is drawn) for every pair. Then, the proportion matrix is made, as in [73] showing the number of times a method is preferred over the others, for all the stimuli. A theoretical example of the type of matrix created is depicted in table 6.1.

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	—	$N_{1>2}$	$N_{1>3}$	$N_{1>4}$	$N_{1>5}$	$N_{1>6}$
S_2	$N_{2>1}$	—	$N_{2>3}$	$N_{2>4}$	$N_{2>5}$	$N_{2>6}$
S_3	$N_{3>1}$	$N_{3>2}$	—	$N_{3>4}$	$N_{3>5}$	$N_{3>6}$
S_4	$N_{4>1}$	$N_{4>2}$	$N_{4>3}$	—	$N_{4>5}$	$N_{4>6}$
S_5	$N_{5>1}$	$N_{5>2}$	$N_{5>3}$	$N_{5>4}$	—	$N_{5>6}$
S_6	$N_{6>1}$	$N_{6>2}$	$N_{6>3}$	$N_{6>4}$	$N_{6>5}$	—

Table 6.1: A general example, showing the number of times each method S at the side is preferred over each one at the top.

Successively, the preference magnitudes are computed, through the maximum likelihood method, having the matrix as input. The process is repeated several times, so as to achieve a Gaussian distribution of the data. In addition, suppose in table 6.2, the columns represent the number of each subset (the second subscript), whereas the rows refer to the six investigated methods (the first subscript). In the last stage, the total preference magnitudes are computed as the average $\bar{x}$, of every row. Yet, before that, the results from every iteration are normalized to the range 0-1.

x_{a1}	x_{a2}	x_{a3}	...	x_{a80}
x_{b1}	x_{b2}	x_{b3}	...	x_{b80}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{f1}	x_{f2}	x_{f3}	...	x_{f80}

Table 6.2: The columns represent the preference score produced from every subset of participants' answers, indicated by the second subscript. The rows refer to the preference for every method in every iteration. The first subscript indicated the type of method i.e method a,b...f.

To compute the confidence intervals for every method, i.e. for every row, which the data are normally distributed in. The following formula for the confidence intervals is used

$$CI = \bar{x} \pm Z_{p/2}\frac{\sigma}{\sqrt{n}} \tag{6.1}$$

where σ is the standard deviation, $Z_{p/2}$ is the student's t inverse cumulative distribution function and n is the number of samples.
However, when estimating the preference magnitudes and for accuracy, the cases where an audio excerpt is always preferred over another are dealt with as follows; a no preference choice is attributed to the method that is never preferred and the counts of preference for the method that is always preferred are not taken into account, while counting the total number of responses.
The values determining the elicited scaling of preference are displayed by bar charts and the confidence intervals by error bars in figure 6.9. Directly the plots show that the method of Avedano-Jot is the most preferred, whereas Faller's method is the least appreciated, followed by Aarts-Irwan method. Although,the magnitudes

clearly differ from each other, the differences among Eq.Levels, PCA and Usher-Benesty methods are somehow marginal. The latter derived results highlight that, the methods having their front cha along with the ambient signals (methods of Faller and Aarts-Irwan) are in general less preferred than methods suggesting the original stereo recording as front channels. The results reliability is corroborated by the fact that the confidence intervals are quite narrow, apart from a small overlap between Eq.Levels and PCA, which occurs the broader error bar of all. It can be deduced that the weaknesses of these algorithms are mainly a combination of heavy processing and the type of stimuli used. For example and looking at figures 6.6 and 6.5, Faller's method is never favourable over others during the trumpet session, yet this doesn't hold in the congas session. As a result, not only the spatial enhancement plays a critical role in the up-mix methods, but the preservation of the front sound image, which also contributes to the optimal spatial quality of a virtual sound environment.

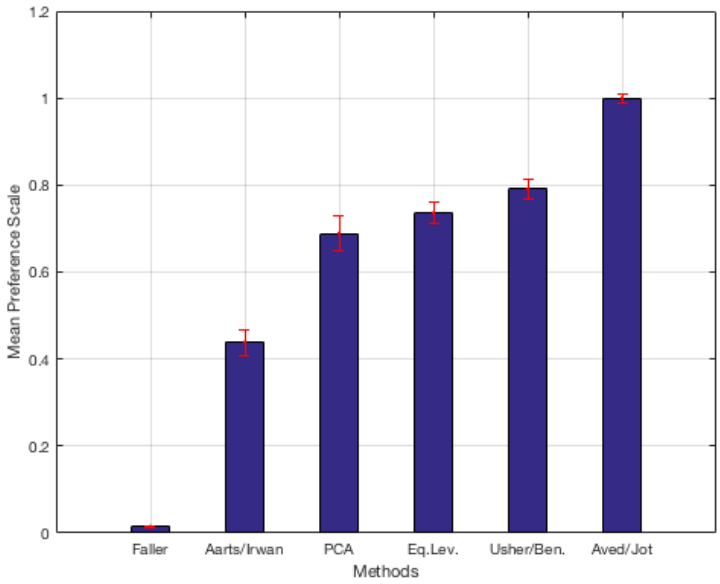


Figure 6.9: Magnitudes of preference, assuming Case V from Thurstone's law of comparative Judgments and bootstrapped 95% confidence intervals

6.3.4 Preference scale per audio stimulus

Apart from the elicited preference scale for every method, the data are further analyzed and categorized by audio stimulus used, in every one of the three sessions of the experiment. Then, Thurstone's scaling is extracted again and depicted in figure 6.10.
This time, more insight is gained in respect to which method is favourable throughout every stage of the listening test and what is the corresponding ranking. Depending on the stimulus, the preference varies for every item. Although high deviations occur for PCA, Aarts/Irwan, and Avedano/Jot, it is stated that, at least for two stimuli, i.e. sessions, either there are no discrepancies at all or they are insignificant. For instance, the levels for the former methods after an abrupt surge, they

occur values between 0.8 and 1, while even though Avedano/Jot's method is the second least appreciated in the voice session, the dip is still more than half of the preference scale. The rest of the systems appear to have a smoother behaviour, with Usher/Benesty's being the smoothest. On one hand, this likely suggests the latter method has a stable rate of preference for every audio material and it's the least affected by the type of stimulus on the other. In general, it can be pointed out that the algorithms have a strong effect, if not a stronger, on the generated preference scale. It is also noted, that in the present case the confidence intervals are not calculated, because the data are sufficient to implement the bootstrapping processing.

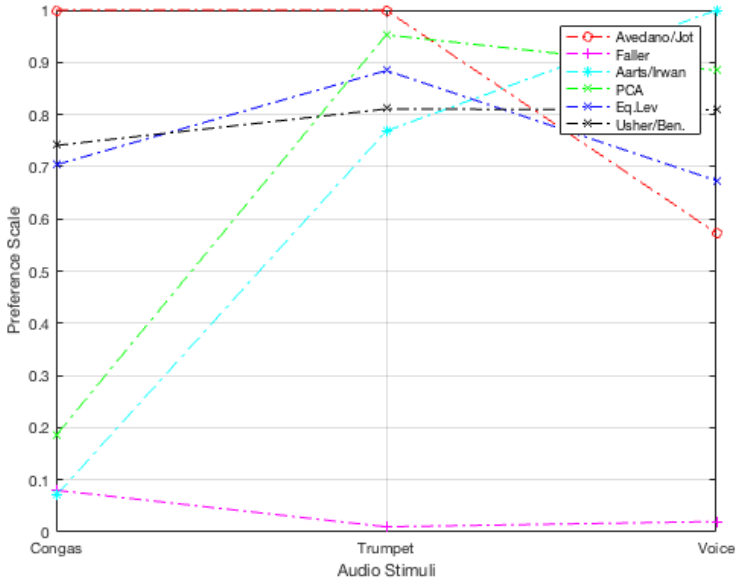


Figure 6.10: Thurstone's Case V preference scale for every type of audio excerpt

6.3.5 Discussion

The results show a clear preference for the Avedano/Jot method, while that of Usher/Benesty is the second most selected. They are followed by Equal Levels and PCA, which score slightly but clearly lower, because their confidence intervals don't appear to overlap with those of the first two methods. On the contrary, methods introducing five channels, like Faller and Aarts/Irwan, scored the lowest preference, which means that increasing the number of channels is not necessarily beneficial for surround sound primary-ambient decomposition methods.
Notwithstanding, taking into account the context simplicity which the experiment is designed upon, the reasons behind the obtained results vary. As it has already been mentioned, some methods are inextricably tied to the features of the stimuli, like for instance PCA, Aarts-Irwan and Faller. For them, the audio content is possibly not suitable hence, it cannot be assumed that these methods will never perform better in another listening test session, using different audio material as input. This fact is asserted, for PCA and Aarts-Irwan methods by looking at figure 6.10. Preference

has been shown to these methods using as stimuli a trumpet and a voice. Both have scored above 0.7 and Aarts-Irwan system is the most preferred for the session where the input is a voice.

Moreover, due to audible undesirable artifacts of certain methods, it is somehow unclear, whether and which degree the listeners' criterium is affected by a forced choice between stimuli bearing unfavourable effects. That is to say, it is ambiguous to which extent the least worst choice determines the listeners' decision. Yet, this is something to be expected, since some up-mixers induce several types of distortion to the audio recordings; depending on the stimuli, Faller's method occur a front image, which sounds rather compressed, lacking in naturalness in the and in Eq. Levels the reverberance image is sometimes unstable, though not always prominent, with the sound being instantaneously hard-panned and consecutively switching between the rear speakers. The same problem is found in PCA for both the front and reverberance imagery, deteriorating the performance of the upmix.

It also happened that some participants informally reported that in a few cases, the front sound source image was more transposed to the right rather in the center. This type of distortion is related to perceptual confusion of the spatial geometry, due to sound features present in both the front and back channels. Even though the rear channels were attenuated by 3 ms, the value it doesn't constitute a perceivable time difference, contributing to a sufficient separation between the front and the rear image. The most appropriate range of values span from 10 ms stated by Rumsey in [63] to 30 ms suggested by Dressler in [64] or even 40 ms reported in [65].

Moreover, neither the decorrelation filter used is adequate to mitigate the described distortion. Another reason could possibly affect the image stability; at least for Usher's method [45], it has also been shown that the ambient image is sometimes unevenly rendered and reproduced by the loudspeakers in the back. It is supposed that the extracted ambience theoretically bears a high degree of diffuseness. Yet, in reality, transients being present in the signals amplify the directivity of the monitors. Hence, the ambient image might be left or right weighted, through radiation of localization cues, which constitute another type of reverberance image distortion. The main artifacts occurred in the listening test for every method are summarized in table 6.3.

In respect to the attributes that affect people's preference, there are several considerable studies investigating the importance of multiple perceptual properties. Not only the spatial characteristics of an upmix system are usually taken into consideration, but also the timbral quality. In [66] it has been shown that it is twice as important as the spatial quality. Successively in the same study, it is inferred that, the quality of the source image channels is more important for experienced audience, however the quality of the rear channels is considered an attribute of greater importance for the inexperienced audience. On the other hand, researches in [67] and [68] excel the importance of attributes related to spatial sound. However, Berg in [69] reports contradictory definitions of the attribute of 'envelopment'. Therefore, it is somehow complex to distinguish among features, which crucially affect the listener preference and the particular level of knowledge was maintained until very lately. In 2017, while the present master thesis work was carried out, a brand new study was released. Francombe et al. in [70] and [71] attempt to define those attributes that

mostly influence a spatial sound listening experience. Their findings primarily introduce the 'amount of distortion' and 'bandwidth' as the most important elements, which discriminate systems with large differences. The former is defined as the total level of distortion and the latter as how broad is the spectral content of the channels. This matches with the subjective results in the present thesis. The method of Avedano-Jot, with minimum or non audible amount of distortion, is clearly differentiated from the other up-mixers, while the amount of audible distortion is mostly associated with the front image produced for certain stimuli by Faller's method. Although, for this system, artifacts in the front channels are reported in [13], their specific type is not mentioned.
'Enveloping' and 'horizontal width' are also found in [70] and [71] to greatly influence the preference. However their influence becomes more relevant, when the quality of evaluated systems is really high, so as they are only corrupted with limited or non audible amount of distortion. Therefore, at least for Avedano-Jot and Usher Benesty, the good level of immersion produced by the sound field and the extended width of the image geometry affects listeners' choice, despite the artifacts being present in the ambience signals. Furthermore, attributes like 'level of reverb', 'phasiness' and 'spectral resonances' are also important, but they usually are properties defined by rather experienced listeners.

Mapping Function (Method of Avedano/Jot)	In a few cases the ambience sounds unnatural
PCA	Severe instabilities in the ambience image. The sound is consecutively being switching between the rear speakers. The same effect in the primary signal, but less pronounced.
Scalar Mask (Eq. Levels of Ambience)	The ambience is consecutively being switching between the rear loudspeakers.
LMS estimates (Faller's Method)	Distortion is occurred in the front image.
NLMS (Method of Usher-Benesty)	The ambience is corrupted by noise at low amplitudes. Transposed front image. Ambience unevenly distributed due to transients.
Time Donain LMS (Method of Aarts-Irwan)	For correlated imputs, either no ambience channels are produced or channels with very low amplitude

Table 6.3: Possible artifacts for the upmix algorithms implemented in the thesis, having as inputs the stimuli used for the subjective evaluation described in the present chapter.

7

Conclusions and Future Work

The present master thesis investigates the major two-to-four or five upmix categories, in the primary-ambience extraction (PAE) framework. The essential aim of the research is an attempt to infer which of the implemented methods is better preferred by the listeners. The basis for each one of them is a masking function (or mapping function), either subjectively tuned-as in Avedano-Jot's method-or analytically obtained by simply assuming equal ambience levels, principal components analysis (PCA), a combination of PCA and a least mean square (LMS) algorithm, a normalized least mean squares algorithm (NMLS) and a method based on least squares estimation (Wiener filtering). The assumption that the primary and ambience are uncorrelated, or vector-wise orthogonal, is common for all methods and it is the basic precondition for extraction algorithms.

Subsequently, the systems are evaluated through objective measures, using as inputs both regular stereo recordings and signals created to bear specific properties, like the uncorrelated stereo impulse and the Gaussian noise. The latter are made in accordance to the specific signal model assumptions, better highlighting the algorithms' advantages and disadvantages. In addition, the primary and ambient components are known beforehand, thereby the decomposition is effectively monitored. In a preliminary context, the methods manage to separate the ambience for every given signal type, however secure conclusion can be drawn, when observing the metrics for the deterministic signals. They suggest that the mapping function (method of Avedano/Jot), the Equal Levels method, the LS estimates (Faller's method) and the NLMS filter (method of Usher-Benesty) have very similar behavior both among them and with the theoretical graphs. An exception could be the fluctuated ambience correlation curves. Yet it is considered a good approximation, since the values are quite low and around zero. Faller's method and PCA yield ambience channels with negative cross-correlation coefficient. On the other hand, it is apparent that the mapping function and the NLMS method insufficiently adapt to signals with sources panned solely to one channel.

Despite the objective metrics offer a useful insight regarding the upmix operation, equivalently important information about the quality of the surround systems is derived by the listeners' feedback through a listening test. Besides, these applications are intended for commercial use. In general, the methods are assessed according to overall preference, using stimuli containing a single source. The findings indicate the mapping function as the most preferred, with NLMS adaptive filter coming next. However, in human voice session the results are narrowly differentiated, in that the Aarts-Irwan method is favored over the others, whereas the mask is the least appreciated, higher only than Faller's method. Hence, in the present upmix scheme

is not possible to completely separate the methods from the input stimuli, a fact which is already been deduced from the objective evaluation as well.

Evaluating the algorithms according to overall preference implies that a set of attributes may have affected the listeners' decision. For instance 'enveloping', 'amount of distortion', 'bandwidth ' and 'naturalness' constitute possible criteria of appreciation for the participants. As a consequence, it is not very easy to point out in specific the attributes which contribute the most to listeners' decision. Nonetheless, recent studies have proved, that there is a trend towards 'enveloping', 'quality of output' and 'horizontal width'. Then, along with the balanced rendering of the ambience channels to the rear loudspeakers, it is equally important to maintain the quality of the front image. Distorted source image is probably the principal key factor for some methods' low performance. All in all, the essential conclusin is that the decisive factor for judging a method over another is the severity of various sorts of artifacts corrupting the channels' quality of every method.

Future Work

Although, the two-to-five channels upmix investigation carried out here covers the major and most promising approaches, some of them actually represent classes of methods rather than unique and absolute implementations without any room for improvement. Hence, some of them like PCA and NLMS adaptive filter are rather general realization of concepts belonging to PAE scheme, with satisfactory computational cost. Therefore, variations of the existing systems could be programmed and tested as well .

Firstly, a version of PCA explained in [12] aims to deal with time-shifted primary components in a stereo channel, a fact commonly observed in commercial audio. This actually violates one of the basic signal assumptions, in that the primary parts are correlated in both channels. In other words they are misinterpreted as ambience. Secondly, a complementary step could be added in the NMLS algorithm as included in [45]. By applying cross-talk, i.e signal leakage from one to another, effective separation is enabled for signals with hard-panned sources. Moreover, a different implementation of the NLMS algorithm could also be done, which is a combination of the method in section 4.6 and the system designed in [72]. It allows for a 'look-ahead' in time of the input signal, so that the filter state will be able better track the optimal solution.

In chapter 5, the ground truth for the algorithms' performance is evaluated based on the cross-correlation coefficient between the extracted and truly embedded ambience. Although, these metrics assess the overall performance of the upmix, they are unable to further analyze the systems, in order to provide with details about the performance degradation. At least for the methods using LMS and PCA, some measures introduced in [48] could also be used. These are the distortion-to-signal ratio (DSR), the leakage-to-signal ratio (LSR) and the interference-to-signal ratio (ISR). The latter quantifies the amount of uncorrelated primary signal extracted from the stereo signal. The first two measure the amount of amplitude scaling of the extracted ambience as compared to the true ambient component and the amount of undesired ambience in the extracted primary (and vice versa), respectively. Con-

sidering that some upmix methods do not extract the primary component, only a limited use of these metrics can be attempted at the moment.
The listening test could be redesigned and improved, yielding information towards a process, that corresponds to a solid product development. For instance, the location of the study should be that of interest, namely the place in which the system would be installed, as the automotive and domestic environment. Regarding the participants, inexperienced listeners should participate to the experiment, because they are the bulk of the future users of the product. In that case, a reliability study should be added to measure the ability of the people to perform the requested task. Here, it is omitted, because most of the participants are students with limited spare time for side activities in an academic environment. Therefore, further extending the duration of the test could actually constitute a problem. In addition, stimuli with more than one sources should be used to test the algorithms, so as to investigate the their effect on different kinds of music.
Taking into account that some methods don't introduce a central loudspeaker in the front, rendering schemes for widening of source image could be used as well. Such techniques are the already mentioned vector based amplitude panning [10] and universal spatial audio coding [74]. A more advance step, would be the real-time upmix of stereo signals, using the equivalent recursive version of the cross-correlation in equation (4.5).

Bibliography

[1] "Multichannel stereophonic sound system with and without accompanying picture",Int.Telecomm. Union, ITU-R BS 775-3,2012

[2] "Advanced Sound Systems for Program Production",Int.Telecomm. Union, ITU-R BS 2051-1,2017

[3] JM Loomis , JR Marston , RG Golledge , RL Klatzky (2005) Personal guidance system for people with visual impairment: a comparison of spatial displays for route guidance, in *J Vis Impair blind, 99(4):219–232*

[4] Airforce Technology (2015) Terma to integrate 3D-Audio/ANR headset in BAE helmets. Available online: http://www.airforce-technology.com/news/newsterma-to-integrate-3d-audioanr-headset-in-bae-helmets-4554218

[5] M. M. Goodwin and J. M. Jot, (2008) Spatial audio scene coding, in *Proc. 125th Audio Eng. Soc. Conv., San Francisco, CA, USA.*

[6] J. Breebaart and C. Faller (2007) Spatial audio processing: MPEG surround and other applications, *Chichester, U.K.: Wiley*

[7] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par (2007) Background, concept, and architecture for the recent MPEG surround standard on multichannel audio compression, in *J. Audio Eng. Soc., vol. 55, no. 5, pp. 331–351, May*

[8] V. Pulkki (2007) Spatial sound reproduction with directional audio coding, in *J. Audio Eng. Soc., vol. 55, no. 6, pp. 503–516, June.*

[9] M. A. Gerzon (1992) General metatheory of auditory localization, in *Proc. 92nd Audio Eng. Soc. Conv., Vienna, Austria.*

[10] V. Pulkki (1997) Virtual sound source positioning using vector base amplitude panning, in *J. Audio Eng. Soc., vol. 45, no. 6, pp. 456–466, June.*

[11] J. Francombe, T. Brookes, R. Mason, R. Flindt, P. Coleman, Q. Liu, and P. Jackson, (2015) Production and Reproduction of Programme Material for a Variety of Spatial Audio Formats, presented at the *138th Convention of the Audio Engineering Society (2015 May), eBrief 199.*

[12] J. He (2017) Spatial Audio Reproduction with Primary Ambient Extraction *Springer (SpringerBriefs in Electrical and Computer Engineering)*

[13] C. Faller (2006) Multiple Loudspeaker Playback of Stereo Signals, in *J. Audio Eng. Soc., Vol. 54, No. 11, 2006 November.*

[14] S. Gajjar (1998) A 3D stereo sound system, in *IEEE Colloquium on Audio and Music Technology: The Challenge of Creative DSP, London, pp. 15/1-15/7.*

[15] C. Faller and F. Baumgarte (2002) Binaural cue coding applied to stereo and multichannel audio compression, in *Proceedings of the Audio Engineering Society 112th Convention, Preprint 5574, Munich, Germany, May 10-13.*

[16] R. Orban (1970) A rational technique for synthesizing pseudo-stereo from monophonic sources, in textiitJournal of the Audio Engineering Society, 18(2):157–164

[17] C. Avedano, J.-M. Jot (2002) Ambience extraction and synthesis from stereo signal for multichannel audio upmix, in *Proc.IEEE Int.Conf. Acoust., Speech, Signal Process.*

[18] M. R. Schroeder (1987) Statistical parameters of the frequency response curves of large rooms, in *J. Audio Eng. Soc., vol. 35, no. 5, pp. 299–305*

[19] J. Merimaa, M. Goodwin J. M. Jot (2007) Correlation-Based Ambience Extraction from Stereo Recordings. *Convention Paper 7282. Audio Engineering Society.*

[20] R. Irwan and R. M. Aarts (2002) Two-to-five channel sound processing, in *J. Audio Eng. Soc., vol. 50, no. 11, pp. 914–926, Nov. 2002.*

[21] M. Goodwin (2008) Geometric Signal Decomposition for Spatial Audio Enhancenment, in *Proc. ICASSP, Las Vegas, NV, USA, pp. 409–412..*

[22] M. Goodwin, J.-M. Jot (2007) Primary-Ambient Decomposition and Vector-Based Localization for Spatial Audio Coding and Enhancement, in *Proc. ICASSP, Honolulu HI, USA, pp. 9-12.*

[23] I. Jolliffe (2002) Principal component analysis, *2nd ed. ed. New York, NY, USA: Springer-Verlag*

[24] J. Usher and J. Benesty (2007) Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer, in *IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 7, pp. 2141–2150, September*

[25] J. Blauert (1997) Spatial Hearing: The Psychophysics of Human Sound Localization. *MIT Press, Cambridge, MA, 1st edition.*

[26] L. Rayleigh (J.W. Strutt) (1907) On our perception of sound direction. *Philosophical Magazine, 13:214–232*

[27] J. Jakka (2005) Binaural to Multichannel Audio Upmix, *Master Thesis, Helsinki University of Technology*

[28] W. Mills (1972) Auditory localization, in *Tobias JV (ed): Foundations of modern auditory theory. Academic Press, New York.*

[29] H. Wallach (1949) On sound localization. *Journal of the Acoustical Society of America, 10:270–274.*

[30] C. Faller and J. Merimaa (2004) Source localization in complex listening situations: Selection of binaural cues based on interaural coherence, in *Journal of the Acoustical Society of America, 116(5):3075–3089.*

[31] B. C. J. Moore (2003) *An Introduction to the Psychology of Hearing. Academic Press, San Diego, USA, 5th edition.*

[32] A. D. Blumlein (1931) U.K. Patent 394,325. Reprinted in Stereophonic Techniques (AES, New York, 1986).

[33] F. Melchior (2017) Investigations on spatial sound design based on measured room impulse responses (Phd Dissertation), *Delft University of Technology, The Netherlands.*

[34] M. H. Tan, J. Hammond (2007) A non-parametric approach for linear system identification using principal component analysis, in *Mechanical Systems and Signal Processing, vol. 21, pp. 1576–1600.*

[35] D. Otte, K. Fyfe, P. Sas and J. Leuridan (1988) Use of principal component analysis for dominant noise source identification, in *Proceedings of the Institution of Mechanical Engineers, International Conference: Advances in the Control and Refinement of Vehicle Noise, C21/88, pp. 123–132.*

[36] J. Hammond, K. Shin (2008) Fundamentals of Signal Processing for Sound and Vibration Engineers, *Wiley.*

[37] B. B. Bauer, Phasor Analysis of Some Stereophonic Phenomena (1961) in *J. Acoust. Soc. Am., vol. 33, pp. 1536-1539*

[38] R. Dressler (1997) Pro Logic Surround Decoder, Principles of Operation, *Dolby Laboratories.*

[39] P. W. Klipsch (1958) Stereophonic Sound with Two Tracks, Three Channels by Means of a Phantom Circuit (2PH3), in *J. Audio Eng. Soc., vol. 6, p. 118*

[40] J. Karhunen, (1994) Stability of Oja's PCA Subspace Rue, in *Neural Comput., vol 6, pp. 739-747.*

[41] R. M. Aarts, R. Irwan, A. J. E. Jansen (2002) Efficient Tracking of the Cross-Correlation Coefficient, in *IEEE Trans. Speech Audio Process., vol. 10, pp. 391-402.*

[42] B. Bernfeld (1973) Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism, presented at *the 44th Convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts), vol. 21, p. 308, May.*

[43] F. Rumsey (1999) Controlled Subjective Assessments of Two-to-Five-Channel Surround Sound Processing Algorithms, in *J. Audio Eng. Soc.,vol. 47, pp. 563–582(July/Aug).*

[44] H. Kutruff (2009). Room Acoustics, *Spoon Press-fifth edition*

[45] J. S. Usher (2007) Subjective evaluation and electroacoustic theoretical validation of a new approach to audio upmixing (Phd Dissertation) , in *McGill University, library and collections, Montreal*

[46] J. Usher, J. Cooperstock, and W. Woszczyk, (2004) A multi-filter approach to acoustic echo cancelation for teleconferencing, in *Proceedings of the 147th Meeting of the Acoustical Society of America, New York.*

[47] G. Potard (2006) 3D-audio object oriented coding (Phd Dissertation), in *University of Wollongong Thesis Collection.*

[48] J. He, Ee-L. Tan, Woon-S. Gan (2014) Linear Estimation Based Primary-Ambient Extraction for Stereo Audio Signals, in *IEEE/ACM Transactions on audio, speech and language processing, vol. 22, no. 2, February 2014.*

[49] A. Papoulis and S. Pillai, (2002) in *Probability, Random Variables and Stochastic Processes. McGraw-Hill, 4th edition*

[50] S. Haykin (2001) *Adaptive Filter Theory. Prentice Hall, Englewood Cliffs, N. J., 4th edition*

[51] J. Proakis and D. Manolakis (1996) *Digital signal processing: principles, algorithms and applications. Macmillan, 3rd edition.*

[52] L. Pelkowitz (1981). Frequency domain analysis of wrap-around error in fast convolution algorithms, in *IEEE Trans. Acoustics, Speech and Signal Processing, 29:413–422.*

[53] P.C. W. Sommen, P. J. VanGerwen, H. J. Kotmans, and A. J. E. M. Janssen (1987). Convergence analysis of a frequency-domain adaptive filter with exponential power averaging and generalized window function, in *IEEE Trans. on Circuits and systems, 34(7):788–798.*

[54] Kihong Shin, Joseph K. Hammond (2007) Fundamentals of Signal Processing for Sound and Vibration Engineers, *Wiley.*

[55] "Multichannel surround sound systems and operations ", AES Technical Council, Document AESTD1001.1.01-10

[56] E. Parizet, N.Hamzaoui, G.Sabatie (2005) Comparison of Some Listening Test Methods, in *Acta Acoustica united with Acustica, vol.91,pp.356-364*

[57] L.Thurstone (1927) A Law of Comparative Judgement, in *Psych. Rev., vol. 34 pp.273-286.*

[58] Harris, John W. Stocker, Horst (1998) "Maximum Likelihood Method." in *Handbook of Mathematics and Computational Science. New York: Springer-Verlag, p. 824, 1998.*

[59] K.Tsukida, M.Gupta (2011) "How to Analyze Paired Comparison Data" in *Tech.Rep. UWEETR=2011-0004, University of Washington*

[60] J. Francombe, T. Brookes, R. Mason, J. Woodcock (2017) Evaluation of Spatial Audio Reproduction Methods (Part 2) in *J. Audio Eng. Soc., vol. 65, pp. 198-211*

[61] B. Efron (1979) Bootstrap methods: Another look at the jackknife, in *Ann. Statist., 7 (1), 1-26*

[62] J.P. Guilford (1954) Psychometric Methods, *New York: McGraw-Hill.*

[63] F. Rumsey (1999) Controlled subjective assessments of two-to-five-channel surround sound processing algorithms, in *Journal of the Audio Engineering Society, 47(7/8).*

[64] R. Dressler (2000). Dolby Surround Pro Logic II Decoder. Principles of operation. Dolby Laboratories Information.

[65] H. Wallach, E. Newman, and M. R. Rosenzweig, (1949) The precedence effect in sound localization, in *The American Journal of Psychology, 62, 315–336.*

[66] F. Rumsey, S. Zielinski, R. Kassier, S. Bech (2005) On the Relative Importance of Spatial and Timbral Fidelities, in Judgments of Degraded Multichannel Audio Quality, in *J.Acoust. Soc. Am., vol. 118, pp. 968-976*

[67] F. Rumsey, J. Berg (2001) Verification and Correlation of Attributes Used for Describing the Spatial Quality of Reproduced Sound, presented at *AES 19TH International Conference:Surround Sound-Techinques, Technology and Perception*, conf. paper

[68] C. Guastavino, B. Katz (2004) Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction, in *J.Acoust. Soc. Am., vol. 116, pp. 1105-1115*

[69] J. Berg (2009) The Contrasting and conflicting definitions of envelopement, presented at the *AES 126TH Convention of Audio Engineering Society*, conv. paper 7050

[70] J. Francombe, T.Brookes and R.Mason (2017) Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences *J. Audio Eng. Soc., Vol. 65, No. 3*

[71] J. Francombe, T.Brookes, R.Mason, J. Woodcock (2017) Evaluation of Spatial Audio Reproduction Methods (Part 2): Elicitation of Perceptual Differences *J. Audio Eng. Soc., Vol. 65, No. 3*

[72] K. Gundry (2001) A new active matrix decoder for surround sound, in *Proceedings of the AES 19th international conference, Schloss Elmau, Germany.*

[73] J.P. Guilford (1954) Psychometric methods, *New York : McGraw-Hill*

[74] M.M. Goodwin, J. M. Jot (2007) (2001) Multichannel surround format conversion and generalized upmix, in *AES* 30^{th} *International Audio Conference on Intelligent Audio Environments.*

A

Appendix 1

A.1 Decorrelation Filter

```
%% Decorrelation FIlter based on Potard's recommendation

%% IIR Decorrelation filter(all-pass)

function[wavout,polb,pola,A,B] = DecAllPass(N,wav)% N is filter order-eve

A = rand(N/2,1)*0.9; %
B= (rand(N/2,1)-0.5)*2*pi; % random phase

% make complex numbers
[real,imag] = pol2cart(B,A);
compli = real +1i*imag;
% second part is complex conjugates roots
compli((N/2)+1:N) = conj(compli);
% make denominator polynomial
pola=poly(compli);
% make numerator polynomial
% coefficients in reverse order to get all?pass response polb=pola(length
polb = pola(length(pola):-1:1);
% filter input signal
wavout=filter(polb,pola,wav);
```

A.2 Listening Test Instructions

Thank you very much for your participation!

In the present test you will be asked to indicate **the magnitude of your overall preference** when comparing a pair of audio stimuli("stimulus A" and "stimulus B"). The scale of preference, represented by the slider will range from

Strong preference for A

to

Strong preference for B

Below there are two screenshots of the GUI in two different modes disabled and enabled.

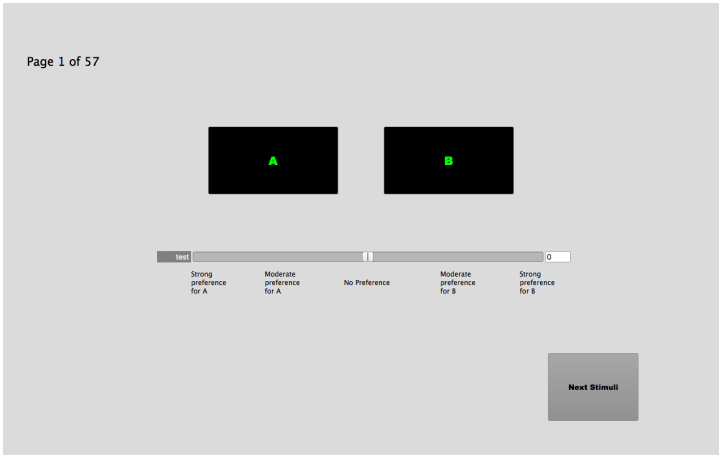


Figure A.1: Disabled GUI

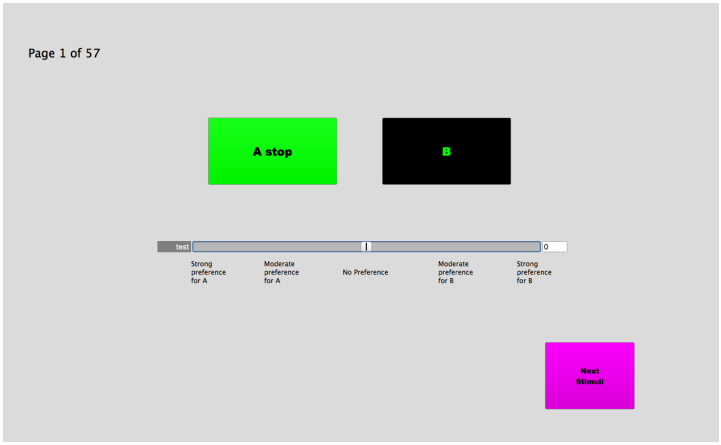


Figure A.2: Enabled GUI

By pressing **'A'** or **'B'** you switch between the two stimuli of the pair. If and only if you listen to each one at least once, a handle will appear on the slider that you can move to indicate your response. Each time, the value of the slider will appear in a white box next to the bar. Once you have given a response, the button **"Next stimuli"** will enable and you will proceed to the next pair
Bear in mind that for some pairs, it might be hard to decide. In these cases, there is also the "No preference".

Always remember: There are no right and no wrong answers!!! It is not tested the ability to perform the task, but the way you perceive the stimuli.

The experiment is split into 3 parts. Before every part a training session takes place, as a familiarization with the interface and to get an idea of the possible range of stimuli. In the "Training Session" your responses are not recorded. Take your time. When the experiment begins the title is changed to "Regular Session".

In a nutshell:

- 3 parts
- 4 pairs of training x 3 training sessions
- 45 pairs of experiment

A.3 The Law of Comparative Judgement (Thurstone's Law)

In Thurstone's model, it is assumed that an option's quality is a Gaussian random variable. The difference between the means of two distributions, during the discriminal processes specifies the difference on the psychological continuum, between the sensations of two stimuli. Standard deviation units are used to quantify the differences between the means of the distributions. If there are several stimuli of similar type being judged according to a certain attribute, then a proportion matrix is obtained like the table created in 6.1, including the percentage of a time a stimulus is selected over another. Therefore, the problem to be solved is to assign a single value on a linear scale. Then, the following equation is introduced, as given in [73] :

$$R_j - R_k = z_{jk}\sqrt{\sigma_i^2 + \sigma_k^2 - 2r_{ik}\sigma_i\sigma_k} \tag{A.1}$$

where R_j and R_k are the mean psychological values assigned to two stimuli, z_jk is the deviation from the mean corresponding with the proportion of occasions a stimulus a is judged over stimulus b, σ_i and σ_j are the standard deviations of distributions R_j and R_k respectively, and r_{jk} is the correlation between the R_j and R_k.

Thurstone's Case V

Thurstone introduced five cases of the method. The most important is case 5, in which it is assumed that the standard deviations are equal and uncorrelated, so

$$R_j - R_k = z_{jk}\sigma_i\sqrt{2} \tag{A.2}$$

Further assuming that the unit of the scale is $\sigma_i\sqrt{2}$, the equation A.3 becomes

$$R_j - R_k = z_{jk}\sigma_i \tag{A.3}$$