PRIMARY-AMBIENT SEPARATION OF AUDIO SIGNALS

A Thesis Presented to the Faculty
of
Nile University

In Partial Fulfillment
of the Requirements for the Degree
of Master of Software Engineering

By
Karim M. Ibrahim
May 2016

Luck Is What Happens When
Preparation Meets Opportunity.
— Seneca

To my parents. . .

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor in Sony Dr. Stefan Uhlich for the continuous support of my thesis and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis.

I would like to thank my supervisor at Nile University Prof. Mahmoud Allam for all his support and encouragement, for pushing me to my limits and giving me the opportunities to always go further while guiding my every step.

My sincere thanks also goes to Sony Stuttgart Technology Center, who provided me an opportunity to join their team as intern, and who gave access to the laboratory and research facilities. Without they precious support it would not be possible to conduct this research.

I thank my fellow labmates both at Nile University and Sony for the stimulating discussions, and for all the fun we have had..

Last but not the least, I would like to thank my family: my parents and my sister for supporting me spiritually throughout writing this thesis and my my life in general.

Stuttgart, January 2016

Abstract

Most audio recordings are in the form of a 2-channel stereo recording. However, many of the new playback sound systems are designed to give a more spatial and surrounding atmosphere that is beyond the content of the stereo recording. Hence, it is an essential step to extract more spatial information from stereo recording in order to reach an enhanced upmixing techniques. One approach of solving the problem is extracting the primary/ambient sources. The problem of primary-ambient extraction (PAE) is one of the challenging problems to decompose a signal into a primary (direct) and ambient (surrounding) sources based on their spatial features. Several approaches have been used to solve the problem, however, there is no clear way to evaluate the different methods of PAE. In this thesis, we propose a new approach to solve the problem based on using trained neural networks as well as modifications to previous methods based on Principal Component Analysis (PCA) to improve their accuracy. An evaluation method is also introduced to provide a way of comparison between different methods. An objective and subjective evaluation is performed to compare between the current state-of-the-art methods and our new proposed methods.

Key words:
Audio Source Separation, Primary/ambient Separation, Spatial Audio, Surrounding Sound Systems, Upmixing.

Contents

List of Figures

1 Introduction

Nowadays, Most available audio recordings are in the form of a 2-channels stereo recording. For a long time, this has been considered sufficient to give the listeners a pleasant experience. However, with new sound systems that can be extended to give a greater sense of a surrounding and enclosing atmosphere, the old recordings fail to support the capabilities of these new systems. Thus, it is important to develop methods of extracting more spatial information from these recordings to enhance the experience of playing them, this process is called upmixing [1, 2]. One approach is by applying audio source separation to extract the original sources from the mixture, which are then rendered for the new playback system [3]. An important distinction between the different audio sources that can be used as a base for separating the sources is the ability to localize the sound sources. Separating sources based on their directional and diffuse features can be used in applications such as upmixing to increase the number of channels and create a surrounding feeling.

5.1 surround systems [4] are an example of a multi-channel sound system commonly used in home theaters that often plays stereo recordings. A practical method of upmixing the stereo sound to the 5.1 system is by separating the primary (localizable) and ambient(non-localizable) sources and play the primary sources on the two front channels to recreate the direct sources as it was intended in the original recording while playing the ambient sources on all channels to give more feeling of the surrounding sound.

Such applications give an increasing importance of developing audio source separation methods. Hence, it has been getting an increasing attention in the research community. Audio source separation can be broadly categorized into two main challenges: blind audio source separation (BASS), where the goal is to extract the different sound sources in the mixture, and primary-ambient extraction (PAE), where the goal is to separate between primary (direct) sources and ambient (diffuse) sources.

The focus of this thesis is on developing new techniques for primary-ambient extraction, improve current methods and make an evaluation of the different state-of-the-art methods and our new proposed methods.

In the remainder of this chapter, we explain the problem definition of audio source separation and primary ambient extraction, the possible application for these techniques, the constraints for an ideal extraction and the notation used in the rest of the thesis.

In chapter 2, we explain the current state-of-the-art methods in extracting the primary and ambient components and state in details of three methods that will be used later on for comparison between our proposed methods and the current ones.

Chapter 3 explains the details of the two proposed methods in this thesis. Chapter 4 shows the evaluation between the proposed methods and the previous methods from the literature using both an objective and subjective evaluation methods. Finally, Chapter 5 concludes the thesis and outlines future work.

1.1 Blind Audio Source Separation (BASS)

Separating sound sources can vary between separating musical instruments from each other, speech from surrounding sources or separating the voices of multiple speakers as in the cocktail party problem [5]. For the general problem of blind audio source separation, the mixture is assumed to be made up of multiple sources as

$$x_i[n] = \sum_{j=1}^{J} a_{ij} s_j[n], \qquad i = 1, 2, ..., I \tag{1.1}$$

where x_i is the i^{th} channel in the mixture out of I channels, s_j is the j^{th} source out of J sources, n denotes the time index and a_{ij} is the scale of contribution of the j^{th} source in the i^{th} mixture channel. The goal is to extract the source estimates $\hat{s}_i$ in a way that minimizes the difference between the estimated and the original source. Figure 1.1 shows the process of extracting the sources in a single-channel mixture.

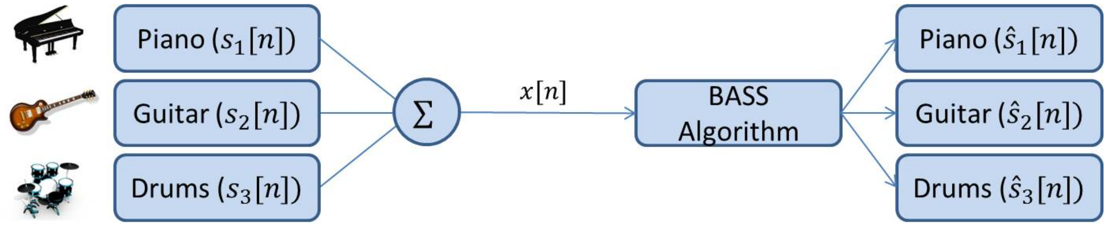


Figure 1.1 – The process of extracting the audio sources from a single-channel mixture

This specific separation problem depends on multiple factors: the number of channels (microphones) used in the recording, the number of sources to be extracted and whether the sources are already known before the extraction or not. In case of not knowing the sources, the problem is called blind source separation and is more challenging than separating specific known sources.

Examples of methods developed to tackle this problem can be reviewed in [6, 7]. However, this problem is concerned more with separating the actual sources from the mixture, while the work presented in this thesis is concerned with a specific kind of separation to capture the spatial characteristics of the recording by separating only the primary from the ambient sources. This is particularly helpful in cases of replicating the recording room and reproducing it through the surrounding sound system to tackle the shortcomings of the typical two-channel stereo system.

1.2 Primary-Ambient Extraction (PAE)

One of the key characteristics in spatial audio is whether an audio source is localizable or not. A localizable source is perceived as coming from a certain direction and the listener can determine this direction, also called primary or directional source. A non-localizable source is perceived as a surrounding sound, coming from all around, also called an ambient or diffuse sound. Ambient sources usually describe the surrounding atmosphere of the recording as shown in figure 1.2. Separating these two types of sources has been receiving an increasing attention for applications such as upmixing [8, 9], multichannel format conversion and headphone reproduction [10, 11].

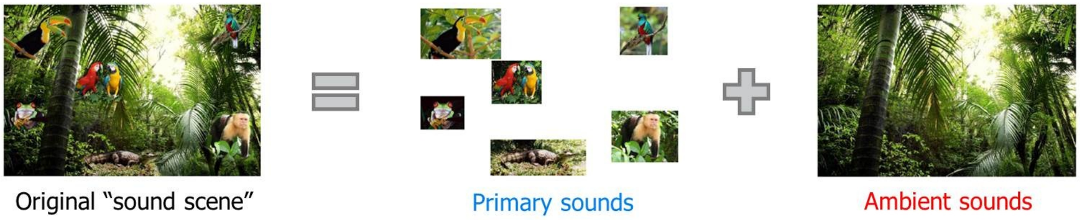


Figure 1.2 – Audio scene of primary and ambient sources

The fact that most of the audio recordings are available in the 2-channel stereo format makes it important to find out methods for upmixing the stereo audio recording to different multichannel systems in a way that improve the surrounding feeling of the audio to match the modern systems. One way to achieve this is separating the audio recording into primary and ambient sources, and to play the different sources on different channels. Therefore, we will discuss here the primary-ambient separation which is a specific BASS problem.

1.2.1 Signal model for PAE

When approaching the problem of primary-ambient extraction, we consider the input signal as a mix of only two kinds of sources; primary and ambient sources. In this thesis, we only approach the problem of separating the mixture of a stereo signal.

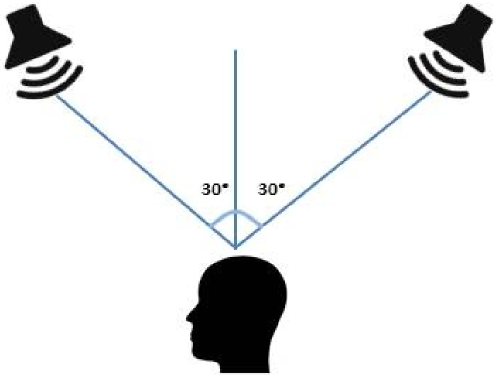


Figure 1.3 – Standard Stereo Loudspeakers Setup

The stereo model is by far the most used playback system and most sound recordings are made for this setup, shown in Figure 1.3. The setup is made of two speakers, hence two channels, on a circle with angles of -30° and 30°. The use of two independent channels is meant to create the illusion of multi-directional sound sources and simulate the sound heard from various directions as in natural hearing.

Since most of the methods of PAE are meant to be applied to stereo recordings, it is necessary to understand the recording process of a stereo mixture. There are two main methods for creating a stereo recording; one is called “true” or “natural” and the second is called “artificial” [12].

The “natural” method consists of capturing an actual live sound with all the natural reverberations and ambiance present by an array of microphones to capture the different locations of the sound sources [13]. A common way of achieving this is by using the binaural recording where two microphones are placed inside a model of a human head to capture the exact signals that reach the human ears as shown in Figure 1.4.

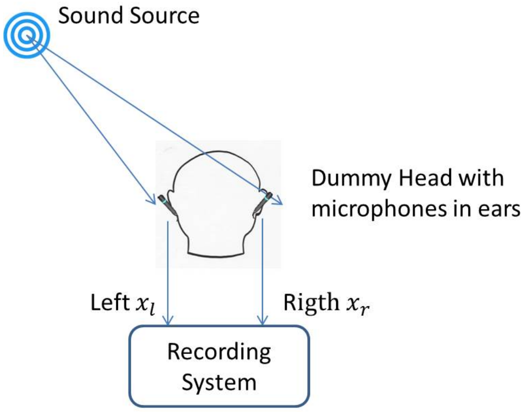


Figure 1.4 – Binaural recording system with in-ear microphones

The “artificial” method is made by reproducing a single channel (mono) sound source over multiple loudspeakers, two in the case of stereo recording, by varying the amplitude of the signal sent to each speaker, also called amplitude panning, and adding a phase shift to create a feeling of the direction of the source [14].

This leads to having two channels that contain both the primary and ambient sources mixed together and the goal is to separate them. The signals can be expressed as follow:

$$x_l[n] = p_l[n] + a_l[n] \tag{1.2a}$$

$$x_r[n] = p_r[n] + a_r[n] \tag{1.2b}$$

where x_l, x_r are the left and right channels of the stereo recording respectively, p_l, p_r are the primary component in each channel, a_l, a_r are the ambient component and n is the time index of the discrete signals.

PAE approaches are better applied in the STFT-domain as it is safer to assume there is only one primary source and one ambient source in each frequency-frame sub-band. The signals are expressed then in the form:

$$X_l[m,k] = P_l[m,k] + A_l[m,k] \tag{1.3}$$

$$X_r[m,k] = P_r[m,k] + A_r[m,k] \tag{1.4}$$

where m, k are the frequency and frame indices respectively.

1.2.2 Sound localization and human auditory system

To be able to precisely separate the primary and ambient sources, it is necessary to understand how the human auditory system works and how it determines the location of a sound source and then use the same characteristics in the separation process.

Localizing sound sources is one important function of the auditory system of all animals that enables them to determine the source of a sound which is a key characteristic of survival [15]. Different animals use different strategies in determining the sound source but almost all of them require the use of two ears and a central processing system that could detect the smallest differences between the two signals detected in the two ears. Our focus here is merely on the human auditory system and the cues it uses in determining the sound source location.

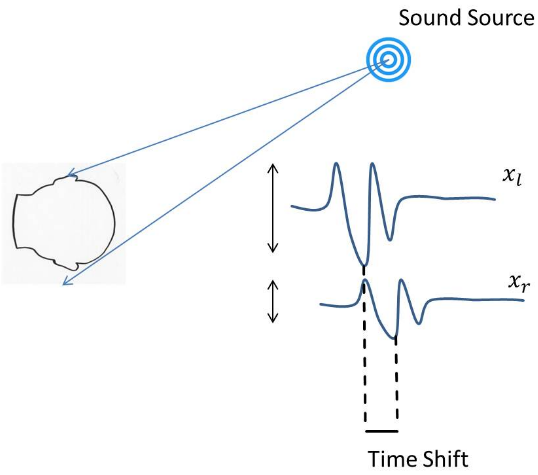


Figure 1.5 – Difference between left and right signals from a non-centered source

The human auditory system uses several cues to localize a sound source, including inter-channel time difference (ICTD), also referred to as inter-aural time difference (ITD), inter-channel level difference (ICLD), also referred to as inter-aural intensity difference (IID), spectral information, time analysis and correlation analysis [16]. The human brain analyzes the two signals from the left and right ear and determines these characteristics to decide the direction of the sound. Two of the basic cues to determine the direction of the source are:

1. Inter-Channel Time Difference (ICTD):
   This refers to the time difference between the sound reaching one ear and the other, a source on the right side of the listener would reach the right ear faster as shown in figure 1.5

2. Inter-Channel Level Difference (ICLD):
   This refers to the difference between the intensity of the signal in one ear and the other. Sound is attenuated as it travels longer distances; hence a source to the right would have higher intensity in the right ear than the left as shown in 1.5

Based on this, a comparison between the two channels should be sufficient to extract the direction information of an audio source. The correlation between the two channels plays a significant role in determining the location of the source, i.e., an ambient source shows no correlation between the two channels, making it impossible for the human auditory system to determine the direction of the sound. Hence, calculating the correlation between the two channels is usually a necessary step in extracting the primary and ambient sources.

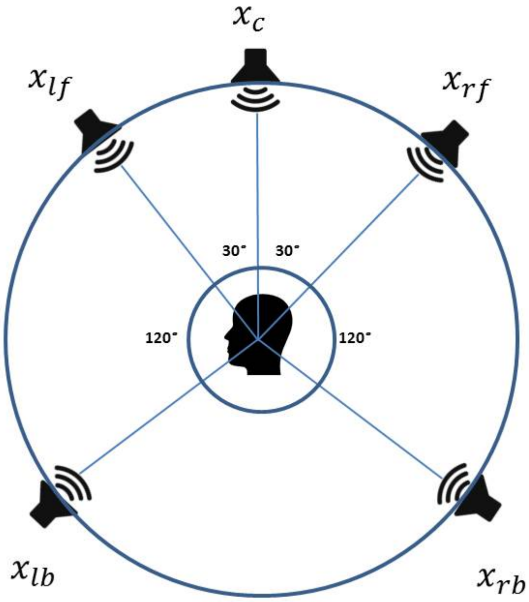


Figure 1.6 – 5.1 surrounding sound system

1.3 Applications of PAE

One important application to primary-ambient separation is the N-to-M channel upmixing, where $M > N$, to be used in sounds systems where the number of loudspeakers is larger than the number of recorded channels. The goal is to enhance the recording in a way that introduces more surrounding feeling and to widen the listening area. However, it is important to preserve the intended stereo image from the original recording. A typical way to achieve this is by separating the primary sources and keep their spatial characteristics to preserve the original image while redistributing the ambient sources to widen and enhance the surrounding feeling.

An example of a common used system that requires upmixing from stereo recordings is the 5.1 system [17]. It is made up of five main channels and one subwoofer channel. The setup of the system is shown in Figure 1.6. To upmix the stereo two-channels recording to the five channels, we need to preserve the original front stage by redistributing the primary sources between the three front channels, while playing the ambient sources in all five channels in a way that preserves the energy level of the original recording.

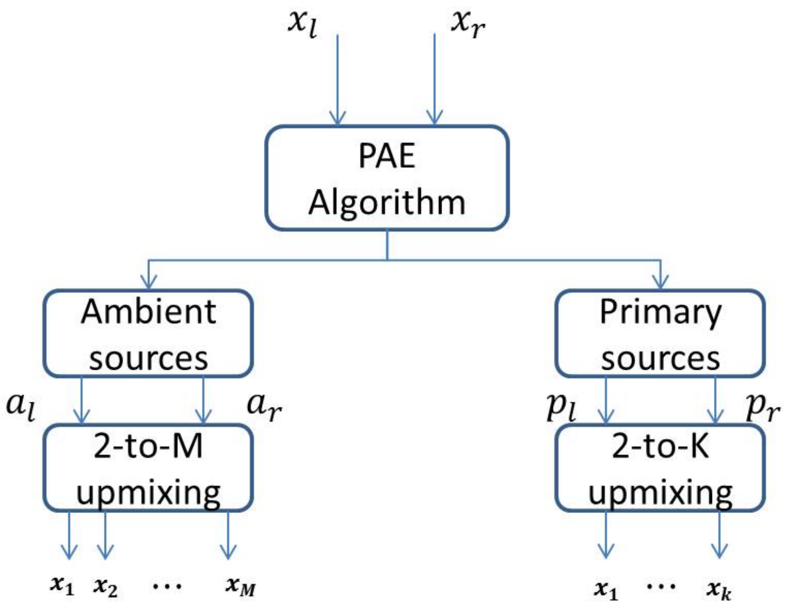


Figure 1.7 – Block diagram of the stereo to 5.1 upmixing using PAE

After separating the primary and ambient sources using one of the PAE methods, the extracted sources are re-panned in a way that the left primary sources p_l are played on the channels x_{lf} and x_c while the right primary sources are played on the channels x_{rf} and x_c and the ambient sources are played all around on the five speakers. Another approach is to play the ambient sources only on the back speakers, however, according to our experiments as shown in the subjective evaluation in section 4.2, it is more appealing to play the ambient sources all around instead of only the back channels. Figure 1.7 shows the block diagram of the upmixing technique from stereo to a generic M channels system that has K front channels. A relevant system is the Quadraphonic system with 4 surrounding channels and 2 front channels where $M = 4, K = 2$. This system will be used later on in Section 4.2 for the subjective evaluation of this thesis.

1.4 PAE assumptions

To accurately separate between the primary and ambient components, we need to define the constraints that achieve the right separation. By definition, the primary sources are localizable while the ambient sources are non-localizable. To find a mathematical representation for this definition, we need to review the sound localizing process in the human auditory system mentioned in section 1.2.2. The key characteristic in localizing the sound sources is the correlation between the two signals reaching the left and right ears. In the case of a complete non-localizable diffuse source, the two signals are expected to be orthogonal in a way that the brain fails to detect any similarity between the left and right signals to extract location information. Similarly, primary sources are expected to be partially or fully correlated. Based on the representation of the stereo signal in equation 1.3 and assuming that the left and right primary components are $\boldsymbol{P}_l, \boldsymbol{P}_r$ respectively, the left and right ambient components are $\boldsymbol{A}_l, \boldsymbol{A}_r$, ω is a scaling factor between the two channels due to ICLD and $\boldsymbol{A}^H$ is the Hermitian transpose of the vector $\boldsymbol{A}$, these constraints are defined as:

1. The primary components are correlated

$$\boldsymbol{P}_l = \omega \boldsymbol{P}_r \tag{1.5}$$

2. The ambient components are orthogonal (fully uncorrelated)

$$\boldsymbol{A}_l^H \boldsymbol{A}_r = 0 \tag{1.6}$$

3. The ambient and primary components are orthogonal to each other

$$\boldsymbol{P}_l^H \boldsymbol{A}_l = 0 \qquad , \qquad \boldsymbol{P}_r^H \boldsymbol{A}_r = 0 \tag{1.7}$$

4. The two ambient components have almost the same energy level

$$\boldsymbol{A}_l^H \boldsymbol{A}_l \approx \boldsymbol{A}_r^H \boldsymbol{A}_r \tag{1.8}$$

Figure 1.8 shows the assumed constraints between the different components. A good example of a recording that shows both ambient and primary mixed sources is the song "Diamonds on the Soles of Her Shoes" by Paul Simon. In this song there is a mixture between the lead singer which is centered and can be perceived as a primary source and a choral in the background that sounds very ambient. An ideal PAE algorithm would separate these two sources in a way that the lead singer is assigned to the primary component and the choral is assigned in the ambient one.

$$\begin{tikzcd}
\boldsymbol{P}_l \arrow[r, leftrightarrow, "\parallel"] \arrow[d, leftrightarrow, "\perp"'] \arrow[dr, leftrightarrow, "\perp"] & \boldsymbol{P}_r \arrow[d, leftrightarrow, "\perp"] \arrow[dl, leftrightarrow] \\
\boldsymbol{A}_l \arrow[r, leftrightarrow, "\perp"] & \boldsymbol{A}_r
\end{tikzcd}$$

Figure 1.8 – Constraints on the primary and ambient components

1.5 Notation

The convention in this thesis is to express signals in the time domain in lower case letter as x, while signals in the STFT domain in upper case letter as X. Scalar variables are expressed in normal italic font as X while column vectors are expressed in bold italic font as $\boldsymbol{X}$ and matrices are expressed in bold non-italic font as $\mathbf{X}$.

Table 1.5 shows the commonly used symbols in this thesis:

$\boldsymbol{x}$	Mixed stereo signal
$\boldsymbol{x}_l$, $\boldsymbol{x}_r$	left and right channels of a sound mixture
$\boldsymbol{p}_l,\boldsymbol{p}_r$	Left and right primary components
$\boldsymbol{a}_l,\boldsymbol{a}_r$	Left and Right ambient components
n	Discrete time index
m	Frequency index
k	Frame index
w_{pl},w_{pr}	weighting factor of the primary source in left and right channels
$\mathbf{C}$	The correlation matrix between the two channels
$\boldsymbol{v}$	The dominant eigenvector of the covariance matrix $\mathbf{C}$

2 Related Work

In this chapter we examine known approaches from literature for solving the PAE problem. PAE is an important step for spatial audio reproduction, 3D audio production and channel upmixing, therefore, many approaches have been proposed to accurately reproduce the ambient and primary sources from the original mix. For how PAE is used in different applications, interested readers can review the applications presented in the literature, such as spatial audio processing in [18, 19, 10] audio mixing [8, 9] and loudspeaker/headphone reproduction systems [11].

Several approaches have been proposed for the PAE problem and some were extended for upmixing applications. Many of these approaches are based on the Principal Component Analysis (PCA) as in [20, 21, 22, 23, 24, 25]. PCA is widely used since the common signal model assumes that the stereo signal is composed of primary sources that are highly correlated and ambient diffuse sources. It is suitable to use a decomposition method such as PCA to extract the correlated primary sources and to assume the ambient sources are the residuals. The work in [22] is also based on the PCA but with an important modification, it takes into consideration the Inter-Channel Time Difference (ICTD) by using a time-shifting technique to improve the extraction of the primary sources. A different approach for the problem is using the least square method to estimate the primary and ambient sources as proposed by Faller in [26] by minimizing the errors between the extracted signals and the original stereo input.

In Avendano's work [27], the approach is to calculate a band-wise inter-channel short-time coherence from the cross- and autocorrelation between the stereo channels which is then used as the basis for the estimation of a panning and ambiance index. In Kraft's approach [28], the proposed method is based on the mid-side decomposition of stereo signals where the two-channels recording is split into "mid" signal that captures the centered content of the recording and a "side" signal that captures the content panned to the left and right side. A method based on separating the ambient sources using an adaptive filter algorithm to detect correlated and uncorrelated signals is proposed in [29].

Though most of the approaches are proposed for stereo recordings, there are approaches aimed at separating the sources in mono single-channels sources. A method based on non-negative matrix factorization (NMF) is described in [30]. A different approach based on supervised learning and low-level features extraction is presented in [31]. A more recent approach was proposed by Uhle in [32] by using adaptive parametric Wiener filters to control to level of interference between the extracted signals.

In the following, we will give more details for the currently commonly used PAE methods. In chapter 4, we will use these methods for comparison to our proposed methods. We picked the PCA-based approach proposed by Goodwin in [20], the approach proposed by Avendano in [27] using inter-channel coherence to calculate panning and ambiance index and the mid-side decomposition method proposed by Kraft and Zölzer in [28].

2.1 Principal Component Analysis (PCA)

Using PCA, the problem of primary-ambient extraction can be viewed as geometric decomposition of the audio signals represented as vectors as proposed in [20, 21]. The goal is to separate the original stereo signal into primary and ambient components by applying the PCA on the signals in the STFT-domain. The PCA is convenient for this problem since the assumption is that the signal is composed of two orthogonal components and one of them is more prominent than the other. For a specific frequency m at frame k, segments of the stereo signal can be defined as

$$\boldsymbol{X}_l{}^T(m,k) = [X_l(m,k-M)\dots X_l(m,k+M)] \in \mathbb{C}^{2M+1} \tag{2.1}$$

$$\boldsymbol{X}_r{}^T(m,k) = [X_r(m,k-M)\dots X_r(m,k+M)] \in \mathbb{C}^{2M+1} \tag{2.2}$$

In its simplest form, segments of the stereo signals can be expressed as (indices are omitted for brevity):

$$\boldsymbol{X}_l = w_{pl}\boldsymbol{P}_l + w_{al}\boldsymbol{A}_l \tag{2.3}$$

$$\boldsymbol{X}_r = w_{pr}\boldsymbol{P}_r + w_{ar}\boldsymbol{A}_r \tag{2.4}$$

where $\boldsymbol{P}_l$ and $\boldsymbol{P}_r$ are the primary unit vectors, $\boldsymbol{A}_l$ and $\boldsymbol{A}_r$ are the ambiance unit vectors, and w_{pl}, w_{pr}, w_{al}, and w_{ar} describe the level and balance of the components.

Similar to the assumptions mentioned in section 1.4 and according to [20], the assumptions for an ideal primary-ambient decomposition are adjusted to match the modified stereo signal model in equations (2.3) and (2.4):

1. The primary components are fully correlated

$$\boldsymbol{P}_l = \boldsymbol{P}_r \tag{2.5}$$

2. The ambient components are orthogonal (fully uncorrelated)

$$\boldsymbol{A}_l^H \boldsymbol{A}_r = 0 \tag{2.6}$$

3. The ambient and primary components are orthogonal to each other

$$\boldsymbol{P}_l^H \boldsymbol{A}_l = 0 \qquad\qquad \boldsymbol{P}_r^H \boldsymbol{A}_r = 0 \tag{2.7}$$

4. The two ambient components have almost the same energy level

$$\boldsymbol{A}_l^H \boldsymbol{A}_l \approx \boldsymbol{A}_r^H \boldsymbol{A}_r \tag{2.8}$$

The first constraint assumes that there is a single primary source to match this constraints; hence, we apply the decomposition in the STFT-domain since it more likely to match the single source condition in case of a single frame-frequency sub-band.

Vector decomposition

In [20, 21] the vector decomposition is based on using PCA. The key task for this method is determining the unit vectors $\boldsymbol{P}_l, \boldsymbol{P}_r$ which represents the primary source. According to equation (2.5), the two unit vectors are equal and only scaled differently between the two channels by w_{pl}, w_{pr}; this simplifies the problem to only finding one vector $\boldsymbol{P}$. The ambient component is then considered to be the residual of the original signal after taking the primary component out:

$$\boldsymbol{A}_l = \boldsymbol{X}_l - w_{pl}\boldsymbol{P} \tag{2.9}$$

$$\boldsymbol{A}_r = \boldsymbol{X}_r - w_{pr}\boldsymbol{P} \tag{2.10}$$

One way of determining the $\boldsymbol{P}$ vector is by using the first principal component, which by definition meets the first condition. For the sake of completeness, we will discuss the computation of the PCA and the decomposition of the signals.

The computation of PCA starts by computing the covariance matrix between the two channels $\mathbf{C} = \mathbf{X}\mathbf{X}^{\mathbf{H}}$, where $\mathbf{X}^T = [\boldsymbol{X}_l, \boldsymbol{X}_r]$. The eigenvector with the largest eigenvalue corresponds to the first principal component, which is the component with the highest energy and is assigned as the primary component $\boldsymbol{P}$; and hence we are assuming that the primary component has relatively higher energy than the ambient one. This decomposition is applied in the STFT-domain; hence, the matrix is:

$$\mathbf{C}(m,k) = \mathbf{X}(m,k)\mathbf{X}^H(m,k) \tag{2.11}$$

Where m is the frequency index and k is the frame index.

Then we apply the Eigendecomposition to get the dominant eigenvector $\boldsymbol{v}$. We define then the primary component's unit vector as:

$$\boldsymbol{P} = \frac{\boldsymbol{v}}{||\boldsymbol{v}||} \tag{2.12}$$

The weights w_{pl}, w_{pr} are defined as the projection of the original channels on $\boldsymbol{P}$:

$$w_{pl} = \boldsymbol{X}_l^H \boldsymbol{P} \qquad , \qquad w_{pr} = \boldsymbol{X}_r^H \boldsymbol{P} \tag{2.13}$$

At this point we can extract all the primary and ambient components in a way that satisfies the first and the third assumptions, but not the second. The two primary components are indeed fully correlated since they are both a scaled version of the unit vector $\boldsymbol{P}$, the ambient components are orthogonal to the primary component as explained in Figure 2.1. However, the two ambient components are not orthogonal to each other, on the contrary, they are fully correlated. This is a main drawback of this method because it results in producing phantom sources of the ambient sources that can be located, which is opposed to the definition of the ambient sources.

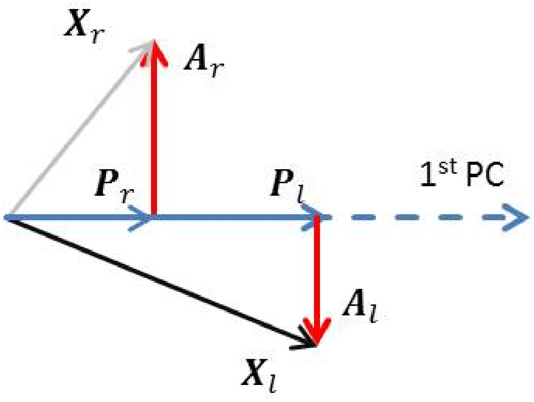


Figure 2.1 – Output components from the PCA method

Another major drawback of this method is the assumption that there is always a primary source in each frequency-frame sub-band which is always prominent; as it is always the first principal component. In case of absence of primary sources, the method would still assign the first principal component, the one with the highest energy, to the primary source, which clearly produces a significant error in this particular case.

One solution for this error is proposed by Goodwin in [20, 33]. When $\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{v} < \theta$, where θ is a threshold defined to limit the leakage of ambient sources in the extracted primary, i.e. there is no reliable principal component, then we get $\boldsymbol{P} = 0$. This is based on the correlation between the two channels which means that the computed ambient will be equal to the signal before decomposition. This results in solving the problem of the absent or weakly-present primary component.

2.2 Inter-channel Coherence Method

In [27], a method based on the inter-channel coherence in the STFT-domain is proposed to extract the primary and ambient components of a mixed signal. Besides the PAE, it introduces a method to estimate the panning coefficients of different sources and how to use them to up-mix and re-pan the sources to a different number of channels. In the following we will briefly explain the PAE approach that is based on the inter-channel coherence that will be used in the evaluation later on. For the full explanation of this approach review Avendano's work in [27].

The approach starts by estimating the correlation between the two channels, first a statistical quantity ϕ relating to the correlation between the channels is defined as:

$$\phi_{i,j}(m,k) = E\{X_i(m,k)X_j^*(m,k)\} \tag{2.14}$$

where m is the frequency index, k is the frame index, E is the expectation operator, i, j are the channels' indices and $*$ denoted the complex conjugation. However, this quantity is stationary in the sense it does not take into consideration the change with time. To track the change with time, a forgetting factor λ is introduced

$$\phi_{i,j}(m,k) = (1-\lambda)\phi_{i,j}(m-1,k) + \lambda X_i(m,k)X_j^*(m,k) \tag{2.15}$$

The inter-channel coherence is then defined as

$$\phi(m,k) = \frac{|\phi_{12}(m,k)|}{[\phi_{11}(m,k)\phi_{22}(m,k)]^{\frac{1}{2}}} \tag{2.16}$$

The inter-channel coherence $\phi(m,k)$ is bounded between 0 and 1, where 1 shows high correlation between the two channels and 0 shows weak correlation. Based on this, the quantity ϕ will have higher values in frequency-frame sub-bands with strong primary component and lower values in sub-bands with a dominant ambient source. An ambiance index is then defined to identify regions with strong ambiance presence as the inverse of the previous quantity

$$\Phi(m,k) = 1 - \phi(m,k) \tag{2.17}$$

The extraction of the ambient component is applied in the form of a mask in the STFT domain made up of the ambiance index per each frequency-frame sub-band after applying a nonlinear transformation on it:

$$A_{l,r}(m,k) = X_{l,r}(m,k)\Gamma[\Phi(m,k)] \tag{2.18}$$

where $A_{l,r}$ is the ambient component of the left and right channel respectively. Γ is the nonlinear function applied on the ambiance index. The characteristics of this function should be that it highly attenuates the output when the ambiance index is small, while being close to one for relatively large $\phi(m,k)$. Γ is defined as:

$$\Gamma(\Phi) = \left(\frac{\mu_1 - \mu_0}{2}\right)\tanh\{\sigma\pi(\Phi - \Phi_0)\} + \left(\frac{\mu_1 + \mu_0}{2}\right) \tag{2.19}$$

where Φ_0 is the threshold between primary and ambient for the ambiance index. μ_1, μ_0 define the range of the output between 0 and 1. σ controls the slope of the function. The values should be picked in a way that separates the components accurately while still preventing any artifacts in the process.

This method results in a smooth extraction of the ambient component with negligible artifacts. There is a considerable number of parameters to push the extraction between a sharp and a smooth extraction. The computation is relatively fast and can be efficiently implemented for up-mixing purposes as described in more details in [27].

2.3 Mid-side Decomposition Approach

In [28], the recently proposed method is based on the mid-side decomposition of stereo signals where the two-channels recording is split into a "mid" signal that captures the centered content of the recording, and a "side" signal that captures the content panned to the left and right sides. The mid-side decomposition is computed as

$$\boldsymbol{X}_M = \frac{\boldsymbol{X}_l + \boldsymbol{X}_r}{2} \tag{2.20}$$

$$\boldsymbol{X}_S = \frac{\boldsymbol{X}_l - \boldsymbol{X}_r}{2} \tag{2.21}$$

The signal model for this method is similar to the one from the PCA-based approach as $\boldsymbol{P_l = P_r = P}$ which is the primary source per each frequency-frame sub-band. The ambient components are $\boldsymbol{A_r} = e^{j\phi}.\ \boldsymbol{A_l}$, which assumes that the two ambient components would be out of phase, i.e. decorrelated, when $\phi = \pi$.

To extract the primary and ambient components, the first step is to estimate the panning coefficients:

$$w_{p,l} = \frac{||\boldsymbol{X}_l||}{\sqrt{||\boldsymbol{X}_l||^2 + ||\boldsymbol{X}_r||^2}} \tag{2.22}$$

$$w_{p,r} = \frac{||\boldsymbol{X}_r||}{\sqrt{||\boldsymbol{X}_l||^2 + ||\boldsymbol{X}_r||^2}} \tag{2.23}$$

using these coefficients with the stereophonic law of sines [34], the angle θ of the source is computed as:

$$\frac{w_{p,l} - w_{p,r}}{w_{p,l} + w_{p,r}} = \frac{\sin\theta}{\sin\frac{\theta_o}{2}} \tag{2.24}$$

where θ_o is the angle between the two speakers.

The primary and ambient components are then determined as:

$$\boldsymbol{P} = \frac{\boldsymbol{X}_l e^{j\phi} - \boldsymbol{X}_r}{w_{p,l} e^{j\phi} - w_{p,r}} \tag{2.25}$$

$$\boldsymbol{X}_l = \frac{w_{p,l}\boldsymbol{X}_r - w_{p,r}\boldsymbol{X}_l}{w_{p,l} e^{j\phi} - w_{p,r}} \tag{2.26}$$

$$\boldsymbol{X}_r = e^{j\phi}.\, \boldsymbol{X}_l \tag{2.27}$$

There is no specific way of determining the value of the angle ϕ, so it is considered as a variable input parameter. By testing, it is better to choose the value of ϕ between $[0.5\pi, \pi]$ to get enough spatial depth by having a phase difference between the two ambient components, while avoiding causing an unpleasant phase cancellation. In the special case where $\phi = \pi$ and the panning parameter are equal to one $w_{p,l} = w_{p,r} = 1$, the output is identical to the original mid-side decomposition in equations 2.20,2.21. For a detailed description of this approach, review [28].

3 Proposed Methods

By reviewing the previous methods proposed in the literature, it is clear that different methods show different drawbacks with room for improvement and there is no complete accurate solution for the problem. In this chapter we will discuss two proposed methods to solve the PAE problem, the first is based on improving the commonly used PCA approach and the second is introducing a new approach for the problem based on using trained neural networks.

3.1 Improving The Principal Component Analysis (PCA) Method

As described in Section 2.1, the PCA-based approach has a number of drawbacks that impairs its accuracy. One of the main drawbacks is that it assumes that there is always a primary source present in each frequency-frame sub-band; hence, after decomposing the signals, it assigns the component with the highest energy (the first principal component) to be a primary source even in the case of having only ambient sources. This leads to extracted ambient sources with generally low level of energy and to a considerable error in the case where no primary source exists.

The solution we propose is to add an adaptive weighting to have more energy in the ambient signal. The weighting is based on the relation between the two channels of the signal in a way that supports the ambient extraction by detecting the level of presence of the primary sources. One way to do this is by considering the second dominant Eigenvalue. In the case of having only ambient sources, the second principal component will be relatively large, due to the orthogonality of the ambient components. This is achieved by comparing the Eigenvalues of the covariance matrix between the two channels. In the case of high correlation, the first (dominant) Eigenvalue will be considerably larger than the second Eigenvalue. In that case it is safe to decompose the signal into primary and ambient components. However, in the case of having a more dominant ambient source, the ratio between the first and second Eigenvalues will be relatively small. There are other weighting schemes to balance the primary and ambient extraction, notably the one proposed by Goodwin in [33] as explained in Section 2.1.

The PAE using our weighting scheme is applied as follows:

1. We start with the original 2-channel signals, $x_l[n]$ and $x_r[n]$. We apply the STFT on the signals to get $X_l[m,k]$ and $X_r[m,k]$, where m is the frame index and k is the frequency index. We calculated the STFT using $\frac{3}{4}$ overlapping Hamming windows of Length 4096 samples, corresponding to a duration of 92.8 milliseconds at a sampling frequency of 44.1 kHz.

2. For each frequency-frame bin we define a vector with the STFT values of the M adjacent frames:

3. For each frequency-frame bin we define a vector with the STFT values of the M adjacent frames:

$$\boldsymbol{X}_l(m,k) = \begin{bmatrix} X_l(m-M,k) \\ \vdots \\ X_l(m+M,k) \end{bmatrix} \quad , \quad \boldsymbol{X}_r(m,k) = \begin{bmatrix} X_r(m-M,k) \\ \vdots \\ X_r(m+M,k) \end{bmatrix} \tag{3.1}$$

   For brevity, the (m,k) index is dropped in the following equations.

4. The decomposition is then applied per frame-frequency index to extract the primary and ambient components in each frame-frequency using these two vectors as shown in Figure 3.1

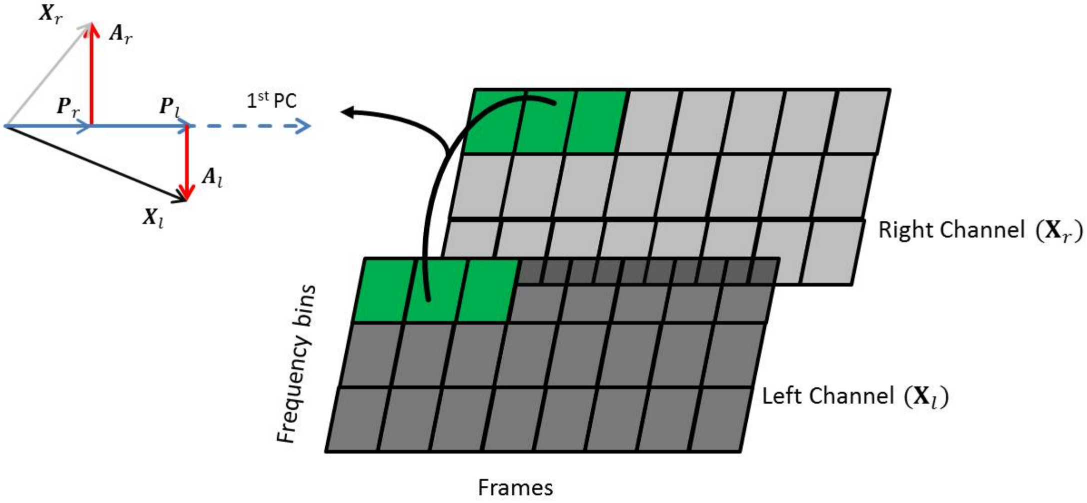


Figure 3.1 – PAE using PCA

5. Next, we calculate the Eigenvalues $\lambda_{1,2}$ and the dominant Eigenvector $\boldsymbol{V}$ of the covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^H$. For simplicity, we can directly calculate the Eigenvalues $\lambda_{1,2}$ from the 2×2 matrix $\mathbf{X}^H\mathbf{X}$ using the simplified formula [35] (for brevity we define

$\rho_{LL} = \boldsymbol{X}_l^H \boldsymbol{X}_l, \rho_{LR} = \boldsymbol{X}_l^H \boldsymbol{X}_r, \rho_{RR} = \boldsymbol{X}_r^H \boldsymbol{X}_r$):

$$\lambda_{1,2} = \frac{1}{2}\left[(\rho_{LL} + \rho_{RR}) \pm \sqrt{(\rho_{LL} + \rho_{RR})^2 - 4(\rho_{LL}\rho_{RR} - |\rho_{LR}|^2)}\right] \tag{3.2}$$

6. Next define a weighting factor ω that corresponds to the level of presence of primary sources in the mixture, i.e. in case of a weak primary source, ω should be small so that the decomposition attenuates the extracted primary component from the PCA accordingly. A starting point is the ratio between the two Eigenvalues $\frac{\lambda_2}{\lambda_1}$. When the correlated primary component is more prominent than the uncorrelated ambient component, the first Eigenvalue would be much larger than the second Eigenvalue and vice versa. Hence, ω was defined initially as :

   $$\omega_{initially} = 1 - \frac{\lambda_2}{\lambda_1} \tag{3.3}$$

   Next, by trying to add more emphasis on the correlation between the two channels in calculating ω, we observed that adjusting the ratio according to the following equation gives better separation

   $$\hat{\lambda}_{1,2} = \frac{1}{2}\left[(\rho_{LL} + \rho_{RR}) \pm \sqrt{(\rho_{LR} + 0.5(\rho_{LL} + \rho_{RR}))^2 - 4(\rho_{LL}\rho_{RR} - |\rho_{LR}|^2)}\right] \tag{3.4}$$

   Hence, the primary weight ω is defined as:

   $$\omega = 1 - \frac{\hat{\lambda}_2}{\hat{\lambda}_1} \tag{3.5}$$

   The stronger the ambient source is the closer the ratio ω gets to 1 and vice versa. Hence, this weighting ensures that in cases of low correlation between the two channels, a higher weight is given to the ambient component, reducing the drawback of dominant primary sources in PCA-based methods.

7. In order to enhance the ambient extraction and limit the leakage of ambient sources into the primary component, we define a threshold θ. The target is to detect the cases where there is no strong presence of a primary source, so all the content is put into the ambient component. The primary component is still weighted by ω in case of passing the threshold to support extracting the ambient component.

   $$\boldsymbol{P_l} = \begin{cases} \omega(\boldsymbol{V^H X_l})\boldsymbol{V} & \omega > \theta \\ \\ 0, & \omega < \theta \end{cases} \qquad \boldsymbol{P_r} = \begin{cases} \omega(\boldsymbol{V^H X_r})\boldsymbol{V} & \omega > \theta \\ \\ 0, & \omega < \theta \end{cases} \tag{3.6}$$

   $$\boldsymbol{A_l} = \boldsymbol{X_l} - \boldsymbol{P_l} \qquad \boldsymbol{A_r} = \boldsymbol{X_r} - \boldsymbol{P_r} \tag{3.7}$$

where $\boldsymbol{P}_l, \boldsymbol{P}_r, \boldsymbol{A}_l, \boldsymbol{A}_r$ are the primary and ambient components of the right and left channel respectively. Figure 3.2 shows an example of the weights extracted from an audio file.

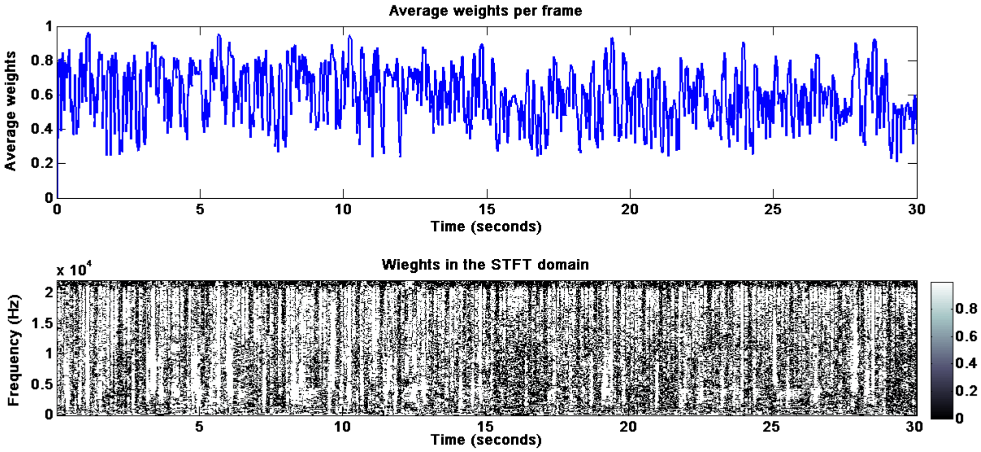


Figure 3.2 – Sample of the weights extracted from an audio file

8. Finally, to extract the information per each frame-frequency, there are two schemes; either to merge the extracted vectors by averaging them or to take out the center point of each vector as shown in figure 3.3. However, the results of both schemes are very similar, so we can use the "take center" scheme to reduce the computation.

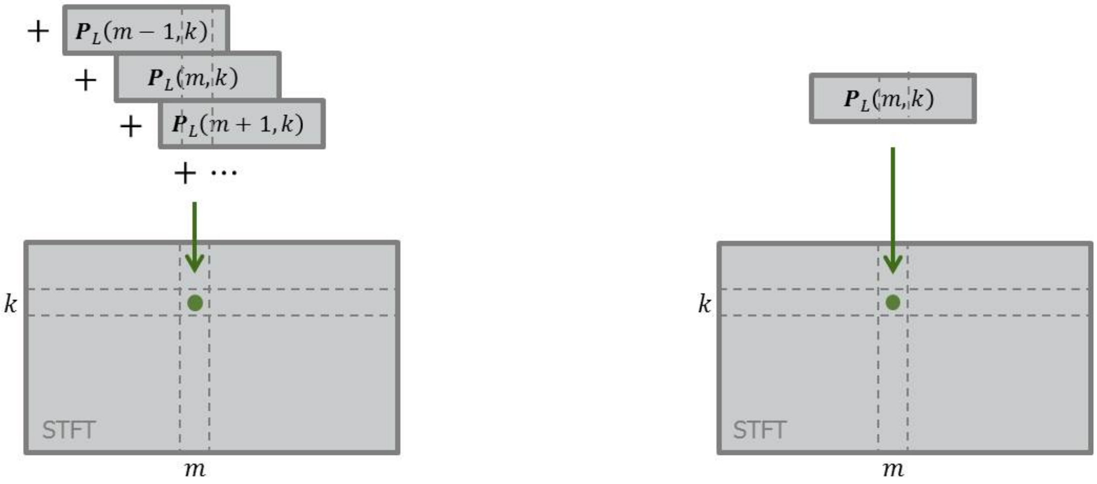


Figure 3.3 – Different schemes of merging the output vectors

Comparison with original PCA approach from [20]

Similar to the original PCA approach, this modified method has the same drawback of not applying the constraint of the orthogonal ambient components. However, it solves the problem of the prominent primary component (the assumption that there is always a primary source), therefore, increases the energy of the extracted ambient signal as shown in Figure 3.4. It also has the advantages of the original PCA of being computationally attractive and suitable for a real-time application.

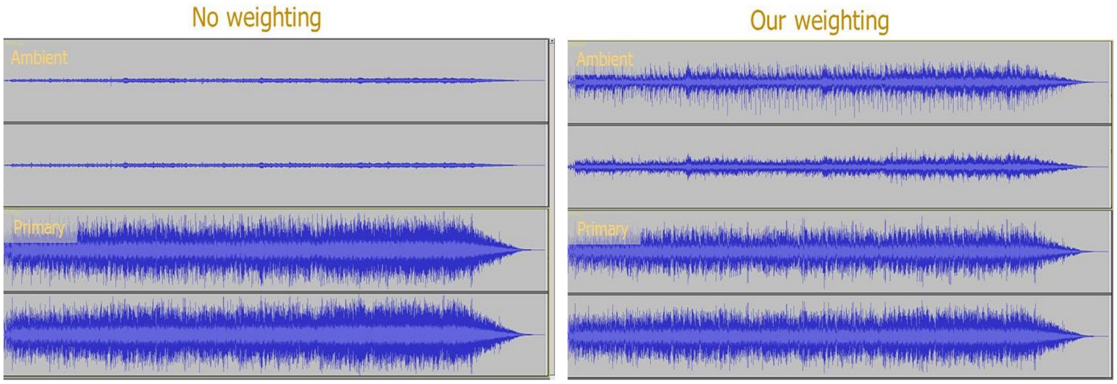


Figure 3.4 – Difference between output signals (ambient top, primary bottom) using PCA with and without the weighting

3.2 Neural Network Method

In this section we will look at the primary/ambient extraction task from a different perspective. We can consider it as a classification problem, where the target is to predict the level of presence of primary and ambient sources in each frequency-frame bin, and then to reconstruct the two signals based on the prediction.

The Setup

In this section we discuss the main steps of setting up the PAE system using a neural network. The three main steps are: collecting a reliable dataset of primary/ambient sources, training the neural network and, finally, applying the classification on the target files.

The Dataset

In order to ensure having a reliable separation, we need to ensure that the data we use for training the neural network is reliable and well-labeled. The separation will be highly dependent on the data we use for training, and for the primary-ambient separation we need to use data that represents the primary and the ambient sources precisely that spans over a large variety of sound sources to ensure that the neural network learns to discriminate between ambient and primary sources of different characteristics.

The dataset we used was made of different types of sources including different music instruments, human voices, animal sounds, surrounding sounds like forests, rain or a cheering crowd. All the sources are labeled to either primary sources that do not include any reverberations or surrounding effects or, conversely, to be ambient. The total number of files used for training is 200 files, 100 of each kind, with a length of 15 seconds for every file.

The next step is to extract the feature vectors from the dataset using the following steps:

1. We clean the data by removing the frames in the STFT domain that contain an energy level less than the average energy level of the input file by 30 dB. This is to remove the frames that have negligible information, as they do not have a large impact on the separation results.

2. The feature vectors are simply the STFT values of each frequency-frame bin combined with two preceding and two succeeding bins for both channels, the length of the feature vector then is 10 by concatenating the frames of the two channels in one vector.

3. Since the STFT values are complex, and complex numbers are not suitable for training the neural network, we then split them into real and imaginary values, ending up with a feature vector of length 20 as shown in figure 3.5

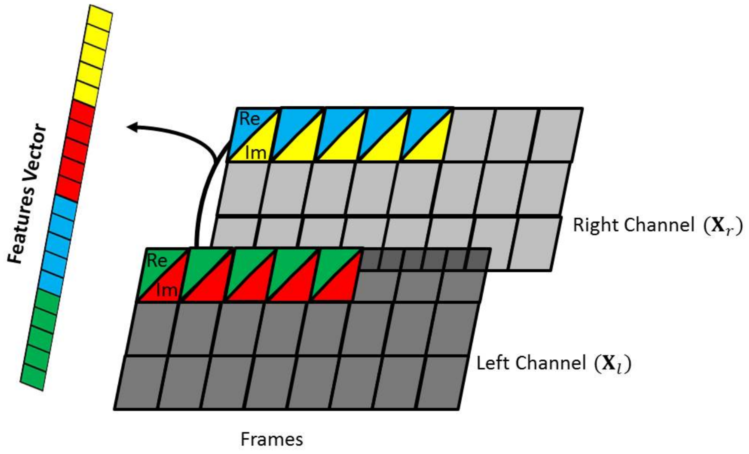


Figure 3.5 – Extracting the feature vectors of the STFT of the input signal

Training the network

The next step is to train a neural network using the data we collected to fit the PAE model. The network is made up of 3 hidden layers, going from 20 features in the input vectors to 15,10 and 2 nodes in the first, second and third hidden layers respectively. The last layer's output range between 1 and 0 and represents the probability of the source being primary. Figure 3.6 shows the steps of extracting and training the neural network, while Figure 3.7 shows the layout of the network.

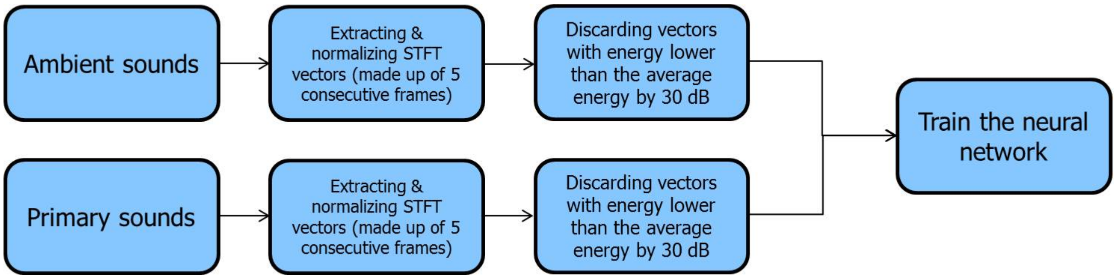

Ambient sounds
Extracting & normalizing STFT vectors (made up of 5 consecutive frames)
Discarding vectors with energy lower than the average energy by 30 dB
Primary sounds
Extracting & normalizing STFT vectors (made up of 5 consecutive frames)
Discarding vectors with energy lower than the average energy by 30 dB
Train the neural network

Figure 3.6 – Steps of extracting the feature vectors and training the neural network

Applying the separation

The final step is to apply the neural network on the target input to be separated to the primary and ambient components. We use the neural network to classify each frequency-frame of the input file to either primary or ambient, then we form a mask of 0's and 1's in the time-frequency domain that corresponds to the classification. Finally, by multiplying the mask to the input STFT we extract the primary component in the time-frequency domain, similarly by applying the complement of the mask we extract the ambient component.Figure 3.8 shows the process of applying the neural network on an input file while Figure 3.9 shows the resulting mask used in extracting the primary component of a test file.

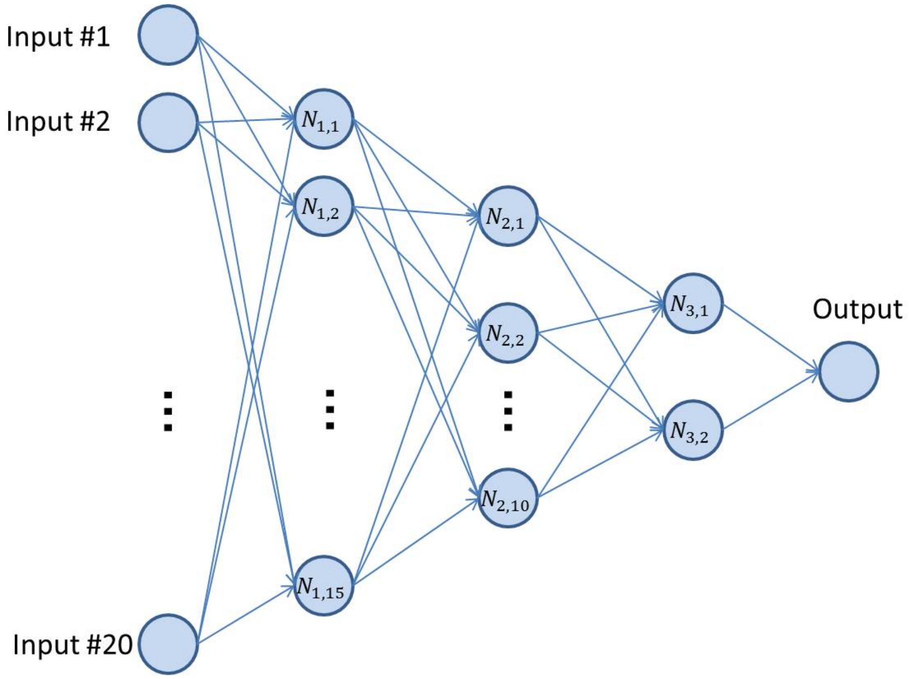


Figure 3.7 – Neural Network Layout

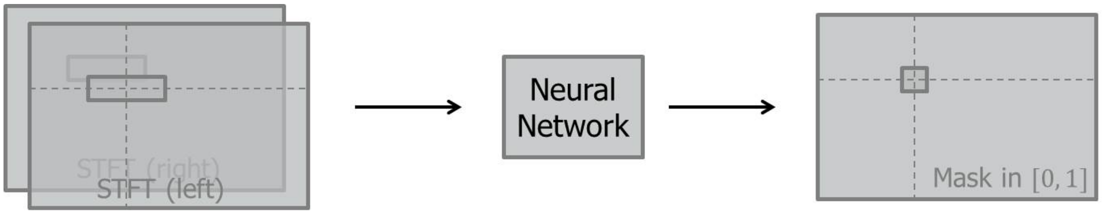


Figure 3.8 – Applying the neural network on an input file

Advantages and further improvement

Once the network has been trained, applying it on target recordings is straightforward and computationally attractive. The network is quite simple and takes relatively short time to process the inputs. An important advantage is the flexibility of changing the focus of the separation and the characteristics of the output. Based on the dataset used to train the network, the output can be shifted towards a specific feature, e.g. training the network using a dataset for ambient sources made only from natural ambient sounds, the output will be more likely shifted towards extracting only the natural ambient sources, leaving out for example the ambient sources such as music instruments. This depends on the application and the target of the separation, however, it is an important feature that can be used differently depending on the case.

Another important feature, the accuracy of the separation is not limited with this particular network layout and training dataset. A more descriptive dataset can be used for the training to guarantee better results. The complexity of the network can be adjusted as well to help improving the separation accuracy.

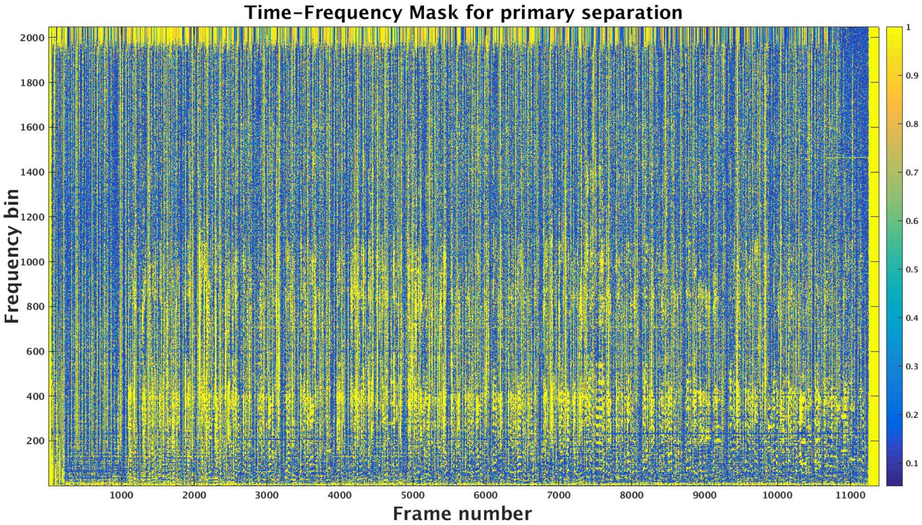


Figure 3.9 – Example of a time-frequency mask for extracting the primary part

4 Evaluation

In this chapter, we discuss the evaluation of different primary-ambient separation methods. There has been several approaches to solve the problem with different extraction methods. However, there is no clear and accurate way of evaluating the performance of these different methods. Hence, here we introduce and perform two evaluation methods to measure the accuracy of the extraction, one is a subjective evaluation methods, based on the user experience, and the second is an objective method based on the performance measurements used for blind source separation described in [36] but adapted for the problem of primary-ambient extraction (PAE).

The evaluation is performed on five different PAE methods:

1. The Primary Component Analysis(PCA) without adding weighting as discussed in section (2.1). Referred to as `PCA`.
2. The modified PCA method by Goodwin in [20, 21]. Referred to as `PCA Goodwin`.
3. The extraction method by Avendano and Jot in [27]. referred to as `Avendano`.
4. The modified PCA method with the weighting scheme proposed by us as described in section (2.2). referred to as the `PCA Improved`.
5. The classification-based approach using neural network proposed by us in section (2.3). Referred to as `Neural Network`.

The subjective evaluation is also performed on different playback systems to rate the actual improvement of using the primary-ambient extraction. The different setups are:

1. Mono one-channel system
2. Stereo two-channels system

3. 4-channel system, stereo played on front speakers and same stereo played on back speakers.

4. 4-channels system, primary played on front speakers and ambient played on back speakers

5. 4-channels system, primary played on front speakers and ambient played on all speakers.

4.1 Objective Evaluation

Our objective evaluation is based on the measurements introduced by Emmanuel Vincent et al. in [36] which is intended to evaluate blind audio source separation (BASS), however, it can be adapted for primary-ambient separation as well. We also introduce an evaluation method based on the error of extracting a source that does not exist in the input as explained in Section 4.1.2. Before discussing how to apply the "BASS" on PAE, we will first describe briefly how it is applied to the general problem of blind source separation.

The problem of BASS is similar to PAE in the sense that there are multiple sources mixed together and the target is to separate this sources. if we denote by $s_j(t)$ the signal emitted by the j-th source out of n sources $(1 \leqslant j \leqslant n)$, $x_i(t)$ the signal recorded by the i-th microphone out of m microphones$(1 \leqslant i \leqslant m)$ and a_{ij} the (causal) source-to-microphone filters, we have $x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) + n_i(t)$, where $n_i(t)$ is some additive sensor noise.

BASS methods are used to separate these sources back from the mix to get the original signals. Ideally the estimated source $\hat{s}_j$, the output of the separation system, would be same as the original source s_j. However, due to limitations of the systems, the estimated and the actual source are not the same. Hence, an evaluation method needs to be used to determine how to judge which system is better in extracting the actual source. The performance measurements introduced in [36] assumes that the true source signal is already known in order to perform the evaluation. The performance measurements are then computed for every estimated source $\hat{s}_j$ by comparing it to the given true source s_j. It decomposes the estimated source as

$$\hat{s}_j = s_j + e_{\text{inter}} + e_{\text{noise}} + e_{\text{artif}} \tag{4.1}$$

where e_{inter}, e_{noise} and e_{artif} are respectively the interference, noise and artifacts error terms. These three error terms should represent the part of $\hat{s}_j$ perceived as coming from the other sources. The second step is to compute the the energy ratios to evaluate the relative amount of each of these terms in the estimated signal. Further details of how the decomposition is performed can be found in [36]

After decomposing the estimated signal, we define numerical performance criteria by computing the energy ratios expressed in decibels (dB). We define Source to Distortion Ratio(SDR)

$$\text{SDR} := 10\log_{10}\frac{||\hat{s}_j||^2}{||e_{\text{inter}} + e_{\text{noise}} + e_{\text{artif}}||^2} \tag{4.2}$$

the Source to Interference Ratio

$$\text{SIR} := 10\log_{10}\frac{||\hat{s}_j||^2}{||e_{\text{inter}}||^2} \tag{4.3}$$

the Source to Artifacts Ratio

$$\text{SAR} := 10\log_{10}\frac{||\hat{s}_j + e_{\text{inter}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2} \tag{4.4}$$

Since the ration varies across time, the perceived quality varies with time as well. These performance measurements are performed locally using a finite length window to compute the errors locally and then averaging to get the errors for the whole signal.

This BASS evaluation method can be adapted to the problem of PAE by considering that the mixture of sources is made out of only two source, one is all ambient s_{ambient} and one is all primary s_{primary}. Ideally the extraction methods should output two sources that are identical to the originals, $\hat{s}_{\text{ambient}} = s_{\text{ambient}}$ and $\hat{s}_{\text{primary}} = s_{\text{primary}}$. However, due to the limitations of the extraction methods, there is interference between the two sources. Hence, e_{inter} indicates the error of interference from the ambient source in the extracted primary source and vice versa, while e_{artif} indicates the error corresponding to the artifacts produced by the PAE separation algorithm.

In this case the SDR for the extracted primary sources can be defined as

$$\text{SDR} := 10\log_{10}\frac{||\hat{s}_{primary}||^2}{||e_{\text{inter}} + e_{\text{artif}}||^2} \tag{4.5}$$

Where $\hat{s}_{primary}$ is the extracted primary source by the algorithm under evaluation, while e_{inter} is an interference caused by the ambient source extracted falsely in the primary component. while e_{artif} is an error of the artifacts caused by the extraction process. Same applies for extracting the SDR for the ambient source.

4.1.1 BASS based evaluation

The evaluation was performed using two databases, one consists of only ambient sources and the second consists of only primary sources. Each dataset consisted of 40 tracks. We used the Matlab toolbox "BSS Eval" [37] for calculating the errors. The evaluation was made out of these main steps:

1. Mixing one ambient source with one primary source after normalizing the two of them.

2. Applying the five different PAE methods to extract the primary and ambient sources.

3. Use the extracted outputs and the original sources to evaluate each method.

4. A baseline is defined by comparing the original ambient or primary sources to the mixture without any separation, i.e. $\hat{s}_{\text{ambient}} = \hat{s}_{\text{primary}} = x$. This is used to define the improvement of each extraction method over the original mixture.

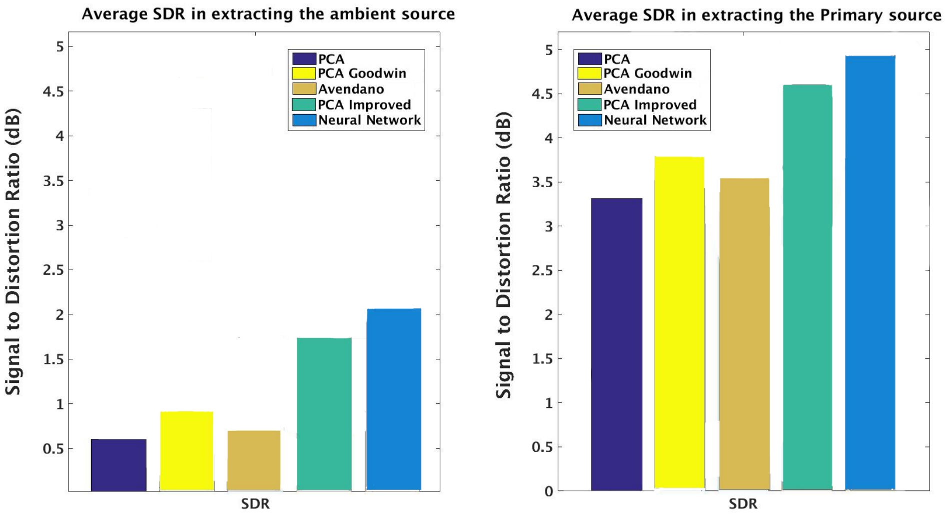


Figure 4.1 – Average SDR in primary and ambient extraction

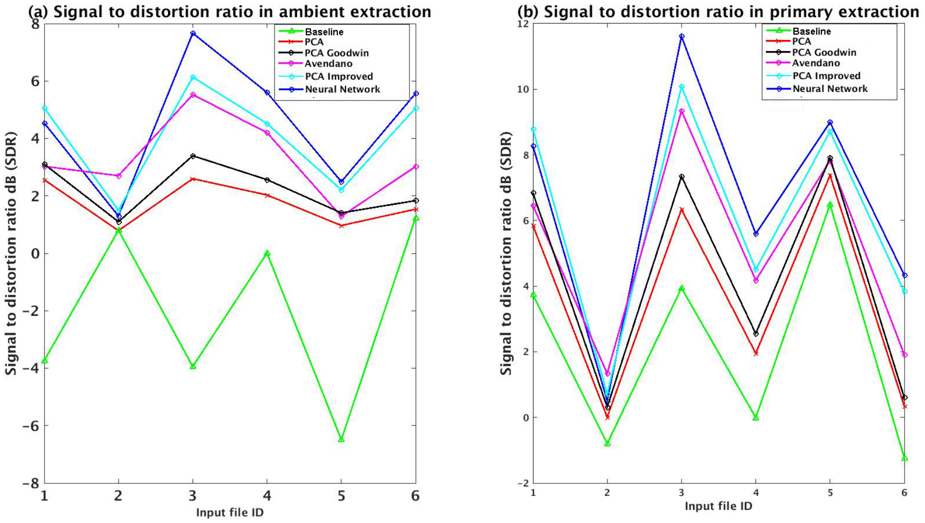


Figure 4.2 – SDR values for a sample of five mixtures

Figure 4.1 shows the average Signal to Distortion ratio (SDR) in extracting both the primary and the ambient sources for different methods. By analyzing the graph, we find that the neural network performs best in terms of SDR in both primary and ambient extraction. Followed by the improved PCA method. This shows how the weighting improves the accuracy of extraction over both the original PCA and the PCA introduced by Goodwin in [20].

Figure 4.2 shows the exact SDR values of a sample of five mixtures with comparison to the baseline in both the primary and ambient extraction. We find that all the methods improve clearly over the baseline without separation. In general the SDR values for the primary extraction is higher than the ambient because the primary sources usually have higher energy.

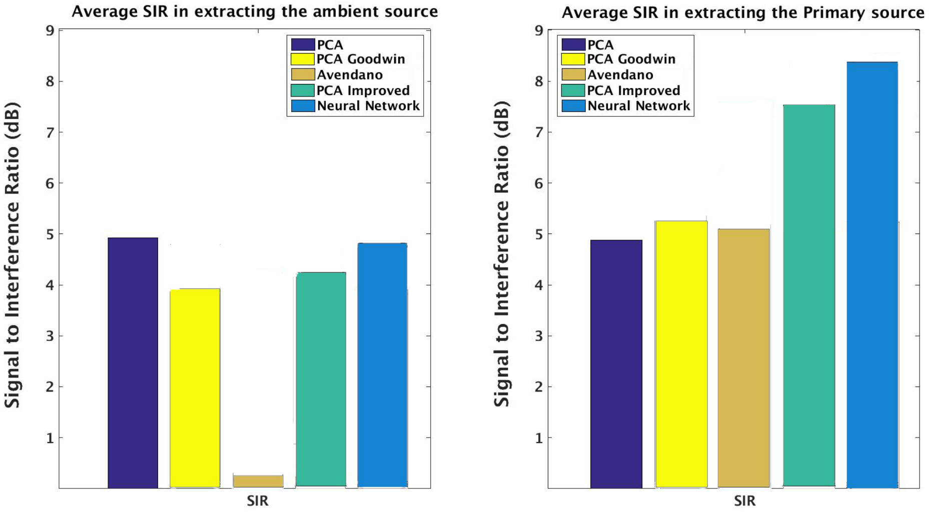


Figure 4.3 – Average SIR in primary and ambient extraction

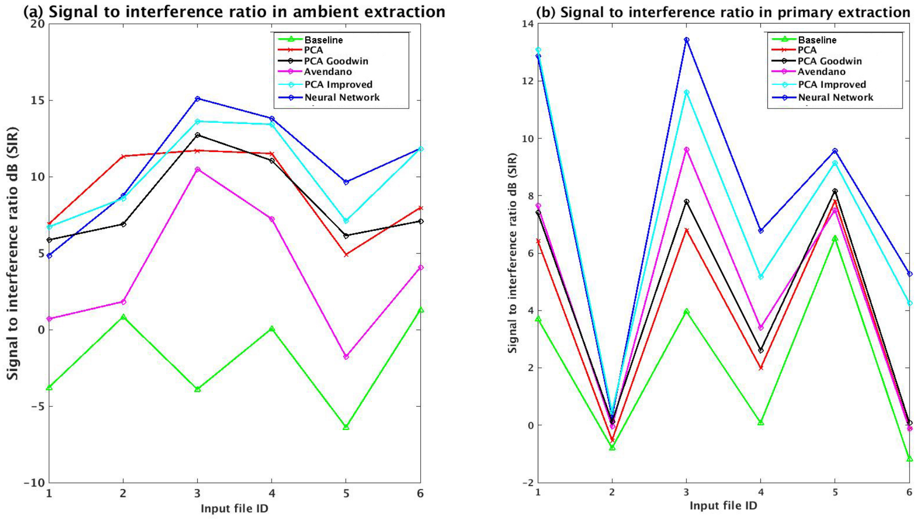


Figure 4.4 – SIR values in a sample of five mixtures

Figure 4.3 shows the average Signal to Interference Ratio. We find that in the ambient extraction the primary interference is minimum in the original PCA, without any weighting, which is expected as the PCA tends to put most of the information in the primary part leaving the ambient with less interference. However, it has the highest interference error in the primary part for the same reason. It is clear that the weighting for the "PCA Goodwin" and our "improved PCA" raises the SIR in the primary part as it tries to put more weight in the ambient part using the weighting discussed in the previous chapters. By skipping the original PCA, due to its deficiency of extracting the ambient part, we find that the neural network still performs best in both cases. Figure 4.4 shows the values of the SIR in the same sample of five mixtures.

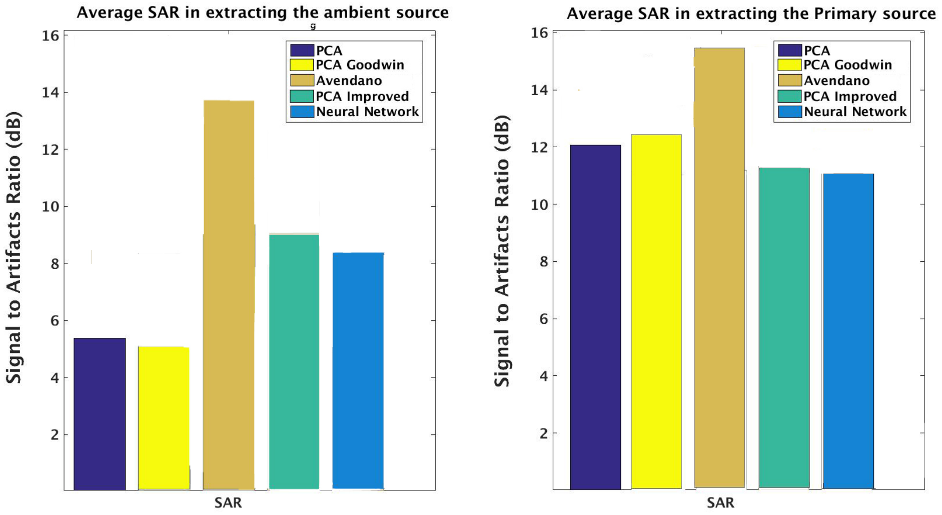


Figure 4.5 – Average SAR in primary and ambient extraction

Figure 4.5 shows the average Signal to Artifacts Ratio. We find that the Avendano's method has the best results in terms of SAR, this is because the method is based on applying a soft mask in the time-frequency plane which results in less artifacts (less musical noise). The neural network is based on classifying every frame-frequency bin to either ambient or primary which would result in producing artifacts. However, by testing the method is a real playback system, this is barely perceptible.

Based on the BASS evaluation, we find that in general the neural network method gives best results followed by the weighted PCA methods. the original PCA and the Avendano's method may give a smoother but less accurate separation.

4.1.2 Single-source-extraction evaluation

The second part of our objective evaluation is based on the error of extracting a source that does not exist. The PAE methods are applied on a signal that contains either all primary or ambient sources. The error is defined by the ratio of the energy of the extracted ambient source to the energy of the original signal, in case of having only primary sources, and vice versa.

Error in extracting primary sources from all-ambient sources

We apply the five different PAE methods on 30 files made from only ambient sources. We then calculate the error as

$$e_{primary} = \sum_{t=0}^{n} \frac{||\hat{s}_{\text{primary}}(t)||^2}{||x(t)||^2} \tag{4.6}$$

where $e_{primary}$ is the error of extracting non-existent primary source, n is the length of the signal, $\hat{s}_{\text{primary}}(t)$ is the extracted primary source, $x(t)$ is the original signal. Ideally, the error should equal zero, since the energy of the extracted primary source should equal zero. However, due to the limitations of the current PAE methods, some parts are incorrectly extracted as primary.

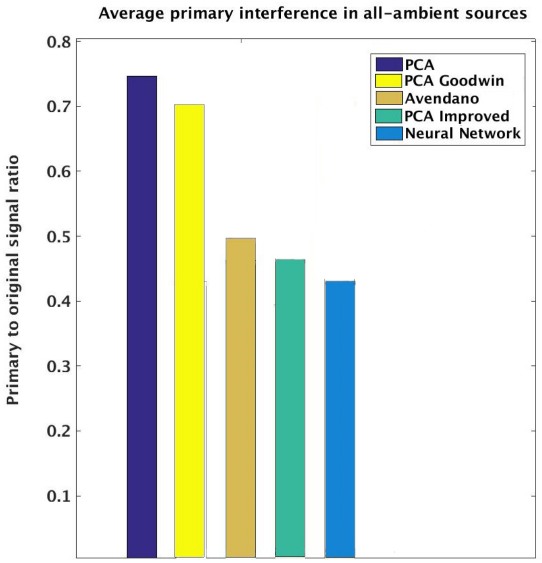


Figure 4.6 – Average primary interference

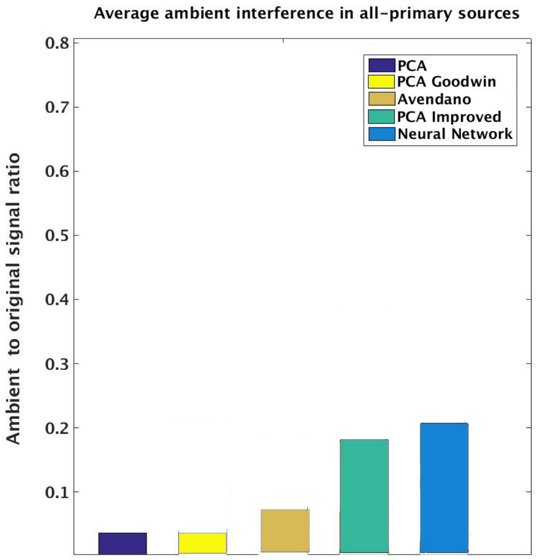


Figure 4.7 – Average ambient interference

Figure 4.6 shows the average error in extracting the primary from all-ambient inputs. Original PCA method performs worst as expected as it tends to put most of the information in the primary part, hence in the case of an all ambient input it performs poorly. However, the weighting introduced by Goodwin and our improved PCA improves the results as it tends to put relatively more information in the ambient part. The neural network still performs best also in terms of this primary extraction error.

Error in extracting ambient sources from all-primary sources

Similarly we apply the five different PAE methods on 30 files made from only primary sources. We then calculate the error as

$$e_{ambient} = \sum_{t=0}^{n} \frac{||\hat{s}_{\text{ambient}}(t)||^2}{||x(t)||^2} \tag{4.7}$$

where $e_{ambient}$ is the error of extracting non-existent ambient source.

Figure 4.7 shows the average error in extracting the ambient from all-primary inputs. Original PCA method naturally performs best as it is better at extracting the primary sources. Only in this case the neural network does not perform best, however the errors are still relatively too small compared to the errors of all-ambient test. This is because it is easier to correctly extract the primary sources than the ambient ones.

4.2 Subjective Evaluation

The second part of the evaluation is based on the user experience. We performed two experiments; the first is to evaluate the different playback systems, and the second is to evaluate the different PAE methods. Both of the experiments were done under the following conditions:

1. The systems were played in a random order
2. The participants did not know previously what system is being played nor did they know what the different systems are.
3. The participants were asked to order the systems from 1 to 5 (where 1 is the preferred system) in terms of:
   (a) Most surrounding sound
   (b) Most natural sound
   (c) overall preference
4. Total number of participants: 11
5. The playback setup was made up from 4 surrounding speaker as shown in figure 4.8
6. For each system two songs (each of length 30 seconds) were played. We selected songs that contain high ambiance and induce a surrounding feeling. The songs are:
   (a) Diamonds on the Soles of Her Shoes by Paul Simon.
   (b) Rock You Gently by Jennifer Warnes.
7. All the systems were adjusted to have the same energy level at the spot where the participant is sitting.

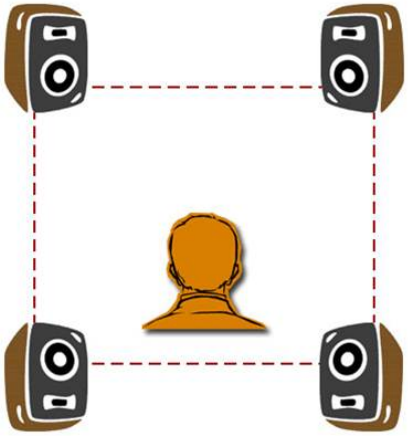


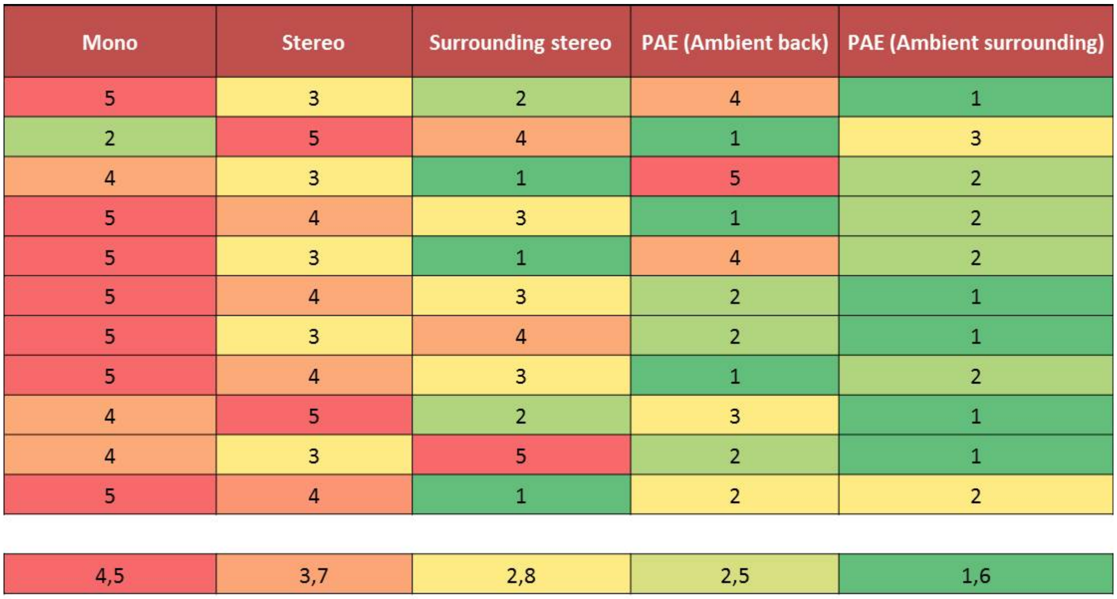
Figure 4.8 – Experiments playback system arrangement

Experiment #1: Different Playback Systems:

The point of this experiment is to evaluate the different arrangements of sound systems and ensure that using the primary-ambient separation results in a better surrounding feeling. The different systems are:

1. Mono single-channel system
2. Stereo two-channels system
3. 4-channel system, stereo played on front speakers and same stereo played on back speakers.
4. 4-channels system, primary played on front speakers and ambient played on back speakers
5. 4-channels system, primary played on front speakers and ambient played on all speakers.

Figure 4.9 shows the ratings of the 11 participants (where 1 is the most favorite and 5 is the least favorite), while the last row represents the average of the ordering. We find that most participants picked the Mono system to be their least favorite as expected, this was acting as an anchor for the experiment to make sure the results are sensible. We find that the stereo and the surrounding stereo (4-channels stereo) is then the least favorite. The primary-ambient separation was picked to be the most preferred system, which is concludes that the separation makes an improvement in the playback systems. The system where the ambient is being played on all speakers is favored over the one where the ambient is played only in the back, this makes sense since the ambient sources should be perceived as coming from all around.



Mono	Stereo	Surrounding stereo	PAE (Ambient back)	PAE (Ambient surrounding)
5	3	2	4	1
2	5	4	1	3
4	3	1	5	2
5	4	3	1	2
5	3	1	4	2
5	4	3	2	1
5	3	4	2	1
5	4	3	1	2
4	5	2	3	1
4	3	5	2	1
5	4	1	2	2
4,5	3,7	2,8	2,5	1,6

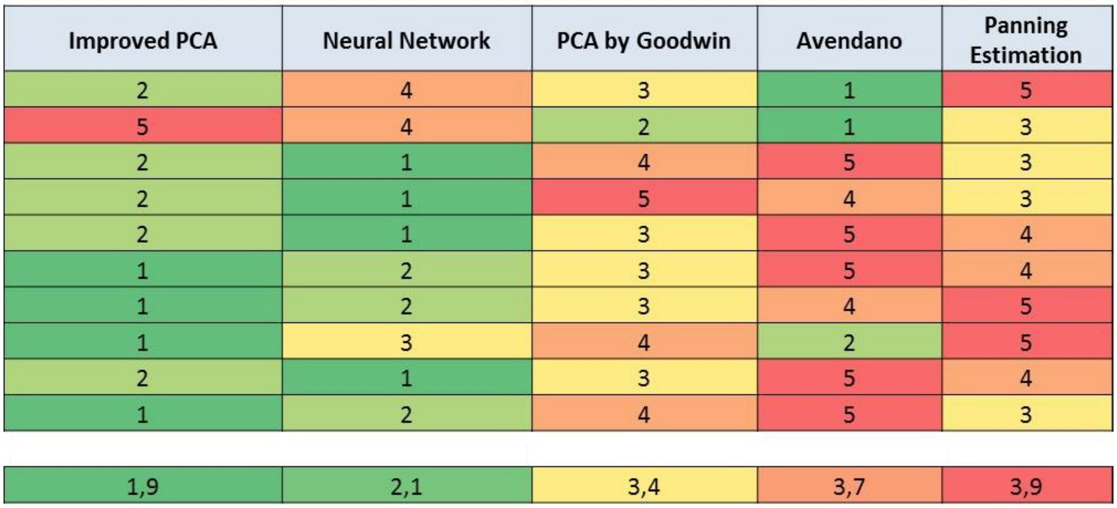
Figure 4.9 – Rating of the different playback systems

Experiment #2: Different Separation Methods:

This experiment was made to evaluate the different separation methods based on the user-experience and to test if the objective evaluation agrees with the actual users' preference. The different PAE methods are:

1. The improved PCA method proposed by us as described in section (2.2).
2. The neural network method proposed by us in section (2.3).
3. The modified PCA method by Goodwin in [20, 21].
4. The extraction method by Avendano in [27].
5. The panning-estimation-based method by Kraft and Zölzer in. [28], Referred to as (Panning estimation method).

Figure 4.10 shows the rating of the participants for the different methods, similar to the previous experiment 1 is the most favorite and 5 is the least favorite. We find that, according to the users' preference, the improved PCA method is the most favorite in terms of being surrounding and appealing. The next favorite is the neural network method, this matches the objective evaluation in selecting these two methods to be the highest rated between all the methods.



Improved PCA	Neural Network	PCA by Goodwin	Avendano	Panning Estimation
2	4	3	1	5
5	4	2	1	3
2	1	4	5	3
2	1	5	4	3
2	1	3	5	4
1	2	3	5	4
1	2	3	4	5
1	3	4	2	5
2	1	3	5	4
1	2	4	5	3
1,9	2,1	3,4	3,7	3,9

Figure 4.10 – Rating of the different PAE methods

Evaluation conclusion:

According to both the subjective and objective evaluation, we find that the neural network and the improved PCA methods perform significantly better than the previously suggested methods. This is perceived in terms of the accuracy of separating the primary and ambient sources and producing an appealing surrounding sound. The subjective evaluation also showed that using the PAE separation improves the sound system and is preferred by the users over the original typical playback systems.

5 Conclusion

Separating the primary and ambient sources from an audio mixture shows much value for applications such as upmixing an audio recording. In this thesis, we explained the need for such a separation and proper ways of using it in upmixing techniques. We presented two new methods of extracting the sources using an adaptive Principal Component Analysis (PCA) to solve the common problem of the dominant primary source. The method shows higher separation quality compared to the classic PCA-based separation methods and other methods from the literature. Additionally we proposed a second method based on using trained neural networks to extract the sources from the STFT representation of the signal.

The proposed weighted PCA method solves one of the drawbacks of the PCA-based methods which is their tendency to put more energy in the primary component even when primary sources are weak. The adaptive weighting tests the level of presence of primary sources and ensures to give a proportional weight to both of the sources based on this estimate. However, this method still shows correlation between the two ambient components leaving room for further improvement in future work.

The neural network approach is rather novel to this problem and acts as an initial prototype for using neural networks in separating the primary and ambient sources. Further work can include investigating using different layouts, different number of hidden layers and other features to get a better and simpler separation.

In the final chapter, we performed two different evaluations to compare the newly proposed methods with selected methods from the literature. The first was an objective evaluation to calculate the levels of interference and artifacts in the extracted sources. According to the results of this evaluation, the newly proposed methods show advancement over the selected methods from the literature. However, the proposed methods are still far from perfect and show much room for further improvement.

The second evaluation was a subjective evaluation to test the users' preferences to different setups and different separation algorithms. According to the evaluation results, using PAE-based surrounding systems gives much better experience for the users and is highly preferred over the traditional playback systems. The evaluation also showed that the users prefer the newly proposed methods over the transitional methods from the literature, indicating that the improvement in the systems is actually observed by the users and enhance their experience.

6 Bibliography

[1] Ville Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. In *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*. Audio Engineering Society, 2006.

[2] Mingsian R Bai and Geng-Yu Shih. Upmixing and downmixing two-channel stereo audio for consumer electronics. *Consumer Electronics, IEEE Transactions on*, 53(3):1011–1019, 2007.

[3] Derry Fitzgerald. Upmixing from mono-a source separation approach. In *Digital Signal Processing (DSP), 2011 17th International Conference on*, pages 1–7. IEEE, 2011.

[4] Xie Bosun. Signal mixing for a 5.1-channel surround sound system'analysis and experiment. *Journal of the Audio Engineering Society*, 49(4):263–274, 2001.

[5] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[6] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):529–540, 2008.

[7] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[8] Mingsian R Bai and Geng-Yu Shih. Upmixing and downmixing two-channel stereo audio for consumer electronics. *Consumer Electronics, IEEE Transactions on*, 53(3):1011–1019, 2007.

[9] Christof Faller and Jeroen Breebaart. Binaural reproduction of stereo signals using upmixing and diffuse rendering. In *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[10] Jeroen Breebaart and Erik Schuijers. Phantom materialization: A novel method to enhance stereo audio reproduction on headphones. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(8):1503–1511, 2008.

[11] Woon-Seng Gan, Ee-Leng Tan, and Sen M Kuo. Audio projection. *Signal Processing Magazine, IEEE*, 28(1):43–57, 2011.

[12] John Eargle. *The Microphone Book: From mono to stereo to surround-a guide to microphone design and application*. CRC Press, 2012.

[13] Günther Theile. Multichannel natural recording based on psychoacoustic principles. In *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[14] Ville Pulkki and Matti Karjalainen. Localization of amplitude-panned virtual sources i: stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–752, 2001.

[15] Arthur N Popper and Richard R Fay. *Sound source localization*. Springer, 2005.

[16] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[17] Geoff Martin, Wieslaw Woszczyk, Jason Corey, and René Quesnel. Sound source localization in a five-channel surround sound reproduction system. In *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.

[18] Francis Rumsey. *Spatial audio*. CRC Press, 2012.

[19] Breebaart Jeroen and Faller Christof. Spatial audio processing: Mpeg surround and other applications, 2007.

[20] Michael M Goodwin and J-M Jot. Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–9. IEEE, 2007.

[21] Michael M Goodwin. Geometric signal decompositions for spatial audio enhancement. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 409–412. IEEE, 2008.

[22] Jianjun He, Ee-Leng Tan, and Woon-Seng Gan. Time-shifted principal component analysis based cue extraction for stereo audio signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 266–270. IEEE, 2013.

[23] Se-Woon Jeon, Dongil Hyun, Jeongil Seo, Young-Cheol Park, and Dae-Hee Youn. Enhancement of principal to ambient energy ratio for pca-based parametric audio coding. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 385–388. IEEE, 2010.

[24] Juha Merimaa, Michael M Goodwin, and Jean-Marc Jot. Correlation-based ambience extraction from stereo recordings. In *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[25] Shi Dong, Ruimin Hu, Weiping Tu, Xiang Zheng, Junjun Jiang, and Song Wang. Enhanced principal component using polar coordinate pca for stereo audio coding. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 628–633. IEEE, 2012.

[26] Christof Faller. Multiple-loudspeaker playback of stereo signals. *Journal of the Audio Engineering Society*, 54(11):1051–1064, 2006.

[27] Carlos Avendano and Jean-Marc Jot. A frequency-domain approach to multichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749, 2004.

[28] Sebastian Kraft and Udo Zölzer. Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain. In *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 2015.

[29] John Usher, Jacob Benesty, et al. Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer. *Ieee Transactions on Audio Speech and Language Processing*, 15(7):2141, 2007.

[30] Christian Uhle, Andreas Walther, Oliver Hellmuth, and Juergen Herre. Ambience separation from mono recordings using non-negative matrix factorization. In *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[31] Christian Uhle and Christian Paul. A supervised learning approach to ambience extraction from mono recordings for blind upmixing. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx08), Espoo, Finland*, pages 137–144, 2008.

[32] Christian Uhle and Emanuel AP Habets. Direct-ambient decomposition using parametric wiener filtering with spatial cue control. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 36–40. IEEE, 2015.

[33] Michael M Goodwin. Adaptive primary-ambient decomposition of audio signals, June 19 2012. US Patent 8,204,237.

[34] Benjamin B Bauer. Phasor analysis of some stereophonic phenomena. *The Journal of the Acoustical Society of America*, 33(11):1536–1539, 1961.

[35] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7:15, 2008.

[36] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.

[37] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent. Bss_eval toolbox user guide–revision 2.0. 2005.