Resynthesis of Acoustic Scenes combining Sound Source Separation and WaveField Synthesis Techniques

Author: Cobos Serrano, Máximo

Director: López Monfort, José Javier

Abstract

Source Separation has been a subject of intense research in many signal processing applications, ranging from speech processing to medical image analysis. Applied to spatial audio systems, it can be used to overcome one fundamental limitation in 3D scene resynthesis: the need of having the independent signals for each source available. Wave-field Synthesis is a spatial sound reproduction system that can synthesize an acoustic field by means of loudspeaker arrays and it is also capable of positioning several sources in space. However, the individual signals corresponding to these sources must be available and this is often a difficult problem. In this work, we propose to use Sound Source Separation techniques in order to obtain different tracks from stereo and mono mixtures. Some separation methods have been implemented and tested, having been one of them developed by the author. Although existing algorithms are far from getting hi-fi quality, subjective tests show how it is not necessary an optimum separation for getting acceptable results in 3D scene reproduction.

Resumen

La Separación de Fuentes ha sido un tema de intensa investigación en muchas aplicaciones de tratamiento de señal, cubriendo desde el procesado de voz al análisis de imágenes biomédicas. Aplicando estas técnicas a los sistemas de reproducción espacial de audio, se puede solucionar una limitación importante en la resíntesis de escenas sonoras 3D: la necesidad de disponer de las señales individuales correspondientes a cada fuente. El sistema Wave-field Synthesis (WFS) puede sintetizar un campo acústico mediante arrays de altavoces, posicionando varias fuentes en el espacio. Sin embargo, conseguir las señales de cada fuente de forma independiente es normalmente un problema. En este trabajo se propone la utilización de distintas técnicas de separación de fuentes sonoras para obtener distintas pistas a partir de grabaciones mono o estéreo. Varios métodos de separación han sido implementados y comprobados, siendo uno de ellos desarrollado por el autor. Aunque los algoritmos existentes están lejos de conseguir una alta calidad, se han realizado tests subjetivos que demuestran cómo no es necesario obtener una separación óptima para conseguir resultados aceptables en la reproducción de escenas 3D.

Author: Cobos Serrano, Máximo, email: macoser1@iteam.upv.es
Director: López Monfort, José Javier, email: jjlopez@dcom.upv.es

Submitting Date: 07-09-07

Contents

Chapter 1

Introduction

Wave-field Synthesis (WFS) is a spatial sound reproduction system that can synthesize a realistic acoustic field in an extended area by means of loudspeaker arrays combined with advanced digital signal processing techniques [3][4]. The sweet spot effect typical of other spatial sound systems as 5.1 surround is eliminated, obtaining more suitable listening areas. In order to recreate an acoustic scene by means of WFS, the different virtual sources (voices, instruments, etc) that compose the scene are positioned in different space locations as shown in Figure 1.1.

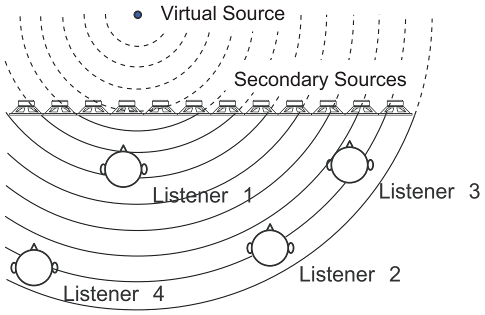


Figure 1.1: Virtual sources can be located anyplace in the space, making all the listeners perceive it with spatial fidelity.

Using the WFS synthesis algorithm the individual excitation signals for each loudspeaker in the array are computed from the individual signals of each instrument in the scene. Therefore, the sound signal of each sound source is needed in the synthesis stage. This presents many advantages because it makes independent the loudspeaker set-up from the sound scene.

However, the requirement of having the separated sound signal for every source can be a drawback in many cases. Despite most of the music is recorded in the studio in separated tracks for each instrument, in the stereo mixdown process this information is lost. Unfortunately, most of the existing recorded material is only available in stereo format, and there is no possibility to obtain the original multitrack recording. On the other hand, source separation has been an intense research field in signal processing during the last years, and a lot of applications are being currently developed, being audio and speech processing two main working disciplines.

In this work we propose to use Sound Source Separation (SSS) techniques to overcome this problem. The scheme in Figure 1.2 proposes the method for obtaining the signals that feed the WFS rendering algorithm. Different separation algorithms have been subjectively tested with different source materials in the 96 loudspeakers WFS array full developed in the GTAC research group. This work is structured as follows: in Chapter 2 an overview of the sound source separation algorithms and applications is given, following in Chapter 3 with a review of the different kinds of algorithms to be tested for monaural one-channel recordings. In Chapter 4 two algorithms for stereo separation are described: the ADRess algorithm and a new algorithm developed by the author. Finally, in Chapter 5 different experiments based on listening tests are carried out and the conclusions obtained from these experiments are presented in Chapter 6.

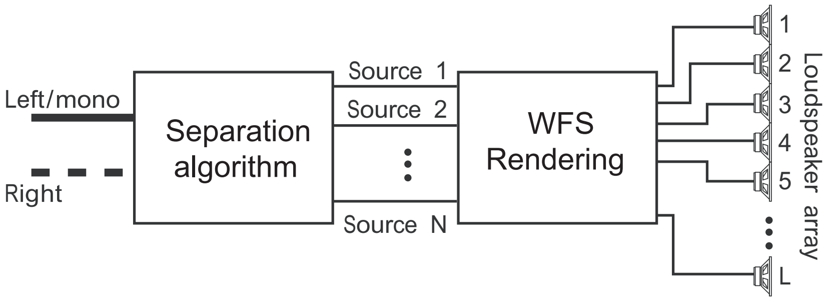


Figure 1.2: Application of SSS to WFS.

Chapter 2

Sound Source Separation

During the last years, Sound Source Separation has been a subject of intense research. When several sound sources are present simultaneously, the acoustic waveform $x(n)$ of the observed time-domain signal is the superposition of the source signals $s_j(t)$:

$$x(t) = \sum_{j=1}^{J} s_j(t), \tag{2.1}$$

where s_j is the j^{th} source signal at time t, and J is the number of sources. Sound Source Separation refers to the task of estimating the signals produced by the individual sound sources from a complex acoustic mixture [9][13][16]. Although human listeners are able to perceptually segregate one sound source from an acoustic mixture, "machine listening" systems are still a great challenge. In some separation techniques, prior information about the sources may be needed. This prior information could be simply the number of sources to be estimated, the kind of sources to be estimated (speech, musical instruments...) or statistical models of the sources. Source separation without prior knowledge of the sources is referred to as *blind source separation*.

2.1 Applications

SSS has a large number of potential applications: high quality separation of musical sources, signal/speech enhancement, multimedia documents indexing, speech recognition in a "cocktail party" environment or source localization for auditory scene analysis. However, current limitations in the existing methods might render some applications if not impossible, at least impractical. A given separation algorithm may perform well on some tasks and poorly on others. It's because of this fact, that depending on the application, various factors affect the difficulty of the separation, and distinct criteria may be used to evaluate the performance of an algorithm. Depending on the application, we could be interested in each individual extracted source or maybe just in extracting one source from the mixture (the target source). For example, extraction of singing voice from a song would be an important achievement [11], not just for remixing purposes, but areas like automatic lyrics recognition, singer identification or music information retrieval.

This work is focused to an Audio Quality Oriented (AQO) application [20]. This means that the extracted sources will be listened to after the separation. In the case of this work, the main purpose will be to examine the possibilities offered by current audio source separation techniques applied to WFS systems. Positioning different sources in different space locations is well accomplished when separated tracks for each source are available. Most of the commercial music productions are recorded this way, but clean information of each source is lost in the mixing process. SSS techniques are the only way to recover the maximum possible information of the different sources. Although separation algorithms produce resulting signals with plenty of artifacts, they might have less importance when separated sources are mixed again in a WFS system. The isolated tracks for each instrument present artifacts that include mainly, inter-source crosstalk and metallic sound. However, when listening to these tracks all together processed with the WFS system, masking mechanisms are involved. This can make the audition of the resynthesized scene perceptually acceptable even if the separation methods applied are not very sophisticated or flawy.

2.2 Traditional Approaches

The main traditional approaches to the source separation problem have always been beamforming and independent component analysis (ICA). Beamforming achieves sound separation by using the principle of spatial filtering. The aim of beamforming is to boost the signal coming from a specific direction by a suitable configuration of a microphone array at the same time that signals coming from other directions are rejected. The array's directivity is directly related to the ratio of its dimension and the wavelength considered. The intrinsic characteristics of audio and speech signals and the acoustic transmission path make difficult the direct application of the beamforming concepts managed in radiocommunications. Basically, the next issues mean a limitation in acoustics beamformers:

- Speech signals have a wide relative bandwidth (5 octaves for speech, 10 octaves in audio). This makes the array's directivity vary a lot in its working frequency range.
- Wavelengths at low frequencies in audio are several meters long, so extremely large arrays would be needed for achieving an acceptable directivity in low frequencies.
- There are a lot of reflections, making the separation even more difficult.
- The array-source distance varies so much relative to the array's length that it's not easy to assume far field conditions in many cases.

In the other hand, Independent Component Analysis models the mixture signal as a standard form of linear superposition of source signals. A instantaneous mixing model of the form $\mathbf{x} = \mathbf{A}\mathbf{s}$ is assumed, where $\mathbf{s}$ is a vector of unknown source signals, $\mathbf{A}$ is a mixing matrix, and $\mathbf{x}$ is a vector of the mixed signals recorded by several sensors. The main assumptions in ICA are that sources involved in the mixing process are statistically independent and non-Gaussian. The separation problem consists in estimating the unmixing matrix (inverse of $\mathbf{A}$). Separation results with ICA are excellent when the assumptions are satisfied, but this not always happens with audio signals

[16]. In addition, the number of sensors should be at least equal to the number of sources to be separated and the difficulty is higher when dealing with convolutive mixtures.

The above techniques are useful just when several observations of the mixture are available. For WFS scene recreation, it would be much more interesting to develop specific algorithms for monaural or stereo recordings. We should concentrate on separation methods where the sources to be separated are not known in advance. These algorithms are based in common properties of real-world sounds, like continuity, sparseness or their harmonic spectral structures.

2.3 One-channel and stereo Sound Source Separation

The first works on one-channel sound source separation concentrated on the separation of speech signals [10][15]. Analysis and processing of music signals have recently received increasing attention [19][23]. Generally speaking, music is more difficult to be separated than speech. Musical instruments have a wide range of sound production mechanisms, and the resulting signals have a wide range of spectral and temporal characteristics. Even though the acoustic signals are produced independently in each source, it is their consonance and interplay which makes up the music [24]. This results in source signals which depend on each other, which may cause some separation criteria, such as statistical independence to fail. Approaches used in one-channel sound source separation which do not use source-specific prior knowledge can be roughly divided into three categories, following the classification proposed in [24]:

- **Model based inference**: These methods use a parametric model of the sources to be separated, and the model parameters are estimated from the observed mixture signal. In music applications, the most commonly used parametric model is the sinusoidal model. The model easily enables the prior information of harmonic spectral structure, which makes it the most suitable for the separation of pitched musical instruments and voiced speech [22].

- **Unsupervised learning**: Unsupervised learning methods apply a simple non-parametric model, and use less prior information of the sources to be estimated. Instead, they try to learn the source characteristics from the observed data. The algorithms can apply information-theoretical principles, such as statistical independence between sources. Algorithms which are used to estimate the sources are based on independent subspace analysis, non-negative matrix factorization [24], and sparse coding [13].

- **Computational Auditory Stream Analysis (CASA)**: CASA methods [25] are based in the ability of humans to perceive and recognize individual sound sources in a mixture referred to as auditory scene analysis [5]. Computational models of this function typically consist of two main stages. First, the mixture signal is decomposed into its elementary time-frequency components. Then, these components are organized and grouped to their respective sound sources. Even though our brain does not resynthesize the acoustic waveforms of each source separately, the human auditory system is a useful reference in the development of one-channel sound source separation systems, since it is the only existing system which can robustly separate sound sources in various circumstances.

Apart from monaural techniques, other approaches have been made to the problem of source separation in music recordings taking advantage of the stereo mixing process [8][1][21][26]. Obviously, if one-channel SSS could be achieved, the stereo problem would be solved just working on each channel independently.

Chapter 3

Monaural SSS Algorithms

In this chapter, two algorithms for monaural sound source separation are described: A Non-negative Matrix Factorization (NMF) algorithm and a speech segregation algorithm. Whereas the NMF algorithm is representative of the unsupervised learning methods, the algorithm for speech segregation is representative of Computational Auditory Stream Analysis (CASA) methods.

3.1 Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria

This algorithm was recently developed by Virtanen [24]. Many unsupervised learning algorithms, for example the standard ICA, require that the number of sensors is larger or equal to the number of sources in order that the separation be possible. By using a suitable signal representation, the methods become applicable with one-channel data.

The most common representation of monaural music signals is based on short-time signal processing, in which the input signal is divided into overlapped frames. Frame sizes between 20 and 100 ms are typical in systems which aim at the separation of musical signals. The representation of the input signal within each frame $t = 1 \ldots T$ is denoted by an observation vector $\mathbf{x}_t$. This observation vector is model as a weighted sum of basis functions $\mathbf{b}_j$, $j = 1 \ldots J$, so that it can be written as

$$\hat{\mathbf{x}}_t = \sum_{j=1}^{J} g_{j,t}\mathbf{b}_j, \quad t = 1, \ldots, T, \tag{3.1}$$

where J is the number of basis functions, and $g_{j,t}$ is the gain of the j^{th} basis function in the t^{th} frame. Note that in 3.1, the subindex j refers to the basis function considered and not to a contributing source as in 2.1. The term *component* refers to one basis function together with its time-varying gain. Each sound source is model as a sum of one or more components, so that the model for source m in frame t is written as

$$\hat{\mathbf{y}}_{m,t} = \sum_{j \in S_m} g_{j,t}\mathbf{b}_j, \tag{3.2}$$

where S_m is the set of components within source m. The sets are disjoint, i.e., each component belongs to one source only. The model in 3.1 can be written in a matrix form as

$$\hat{\mathbf{X}} = \mathbf{B}\mathbf{G}, \tag{3.3}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_J]$ is the *basis matrix*, and $[\mathbf{G}]_{j,t} = g_{j,t}$ is the *gain matrix*.

The spectrograms of musical signals often have a unique decomposition into non-negative components, each of which represents parts of a single sound source. Therefore, in the signal model of 3.3 the element-wise non-negativity of $\mathbf{B}$ and $\mathbf{G}$ alone is a sufficient condition for the separation of sources in many cases, without an explicit assumption of the independence of the sources. This way, the observed magnitude spectrogram $\mathbf{X}$ is model as a product of the basis matrix $\mathbf{B}$ and the gain matrix $\mathbf{G}$, while restricting $\mathbf{B}$ and $\mathbf{G}$ to be entry-wise non-negative.

Estimation of B and G is done by minimizing a cost function $c(\mathbf{B}, \mathbf{G})$, which is a weighted sum of three terms: a reconstruction error term $c_r(\mathbf{B}, \mathbf{G})$, a temporal continuity term $c_t(\mathbf{B}, \mathbf{G})$, and a sparseness term $c_s(\mathbf{G})$:

$$c(\mathbf{B}, \mathbf{G}) = c_r(\mathbf{B}, \mathbf{G}) + \alpha c_t(\mathbf{B}, \mathbf{G}) + \beta c_s(\mathbf{G}), \tag{3.4}$$

where α and β are the weights for the temporal continuity and sparseness term, respectively.

Real-world sounds usually have a temporal structure, and their acoustic characteristics vary slowly as a function of time. The temporal continuity is considered by assigning a cost to large changes between the gains $g_{j,t}$ and $g_{j,t-1}$ in adjacent frames. To prevent the numerical scale of the gains from affecting the cost, the gains are normalized by their standard deviation estimates σ_j, so it can be written as

$$c_t(\mathbf{G}) = \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \sum_{t=1}^{T} (g_{j,t} - g_{j,t-1})^2 \tag{3.5}$$

A signal is said to be sparse when it is zero or nearly zero more than might be expected from its variance. Such a signal has a probability density function or distribution of values with a sharp peak at zero and fat tails. Adding an sparseness objective can improve the quality. The sparsity cost function is

$$c_s(\mathbf{G}) = \sum_{j=1}^{J} \sum_{t=1}^{T} f(g_{j,t}/\sigma_j), \tag{3.6}$$

where $f(\dot{})$ is a function which penalized non-zero gains. Commonly $f(x) = |x|$ is used.

The overall estimation algorithm is the following:

1. Initialize each element of $\mathbf{B}$ and $\mathbf{G}$ with the absolute value of Gaussian noise.

2. Update $\mathbf{B}$ using the multiplicative update rule:

$$\mathbf{B} \leftarrow \mathbf{B}. \times \frac{\frac{\mathbf{X}+\epsilon}{\mathbf{B}\mathbf{G}+\epsilon}\mathbf{G}^T}{\mathbf{1}\mathbf{G}^T} \tag{3.7}$$

3. Update the gains using projected steepest descent:

   - Calculate the gradient of $c(\mathbf{B}, \mathbf{G})$ with respect to $\mathbf{G}$ using 3.4.

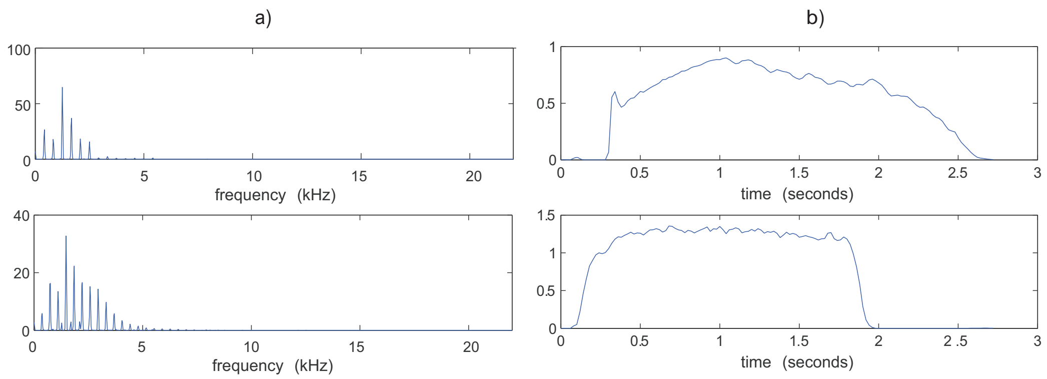


Figure 3.1: NMF components estimated from a mixture signal of two notes (trumpet and oboe). a) Gains. b) Basis functions.

- Update $\mathbf{G} \leftarrow \mathbf{G} - \mu \nabla c(\mathbf{B}, \mathbf{G})$. The positive step size μ is adaptively varied using the bold driver algorithm [ref 150].
- Set negative entries of $\mathbf{G}$ to zero.

4. Evaluate the value of the cost function $c(\mathbf{B}, \mathbf{G})$.

Figure 3.1 shows the result of applying the algorithm to a signal made up of two notes played by two different instruments (trumpet and oboe). Two components were calculated corresponding to the harmonic spectrum of the two notes played. Their temporal gain is showed in the two upper plots and they give information about when the notes are being played and its level over time.

3.2 Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation

The next algorithm is designed specifically for extracting a target speech from a mixture and it is representative of CASA algorithms, recently developed by Hu and Wang [7]. The overall model is a multistage system, as shown in Figure 3.2.

Signal Decomposition

In the first stage, the input mixture (sampled at 16 kHz) passes an auditory filterbank in consecutive time frames, resulting in a decomposition into a two-dimensional time-frequency map. The gammatone filterbank is a standard model of cochelar filtering. The impulse response of a gammatone filter is

$$g(t) = \begin{cases} t^{l-1} \exp(-2\pi bt) \cos(2\pi ft), & t \geq 0 \\ 0 & \text{else} \end{cases} \tag{3.8}$$

where $l = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter. In each filter channel, the otuput is divided into 20 ms time frames with

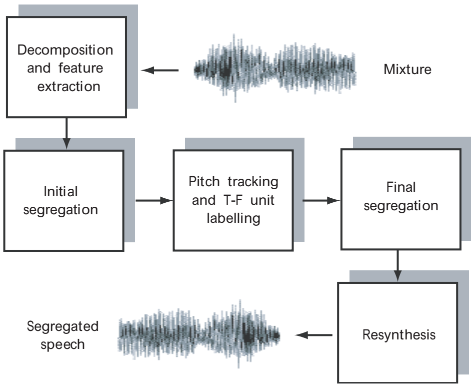

Decomposition and feature extraction
Mixture
Initial segregation
Pitch tracking and T-F unit labelling
Final segregation
Resynthesis
Segregated speech

Figure 3.2: Schematic diagram of the CASA multistage system for speech segregation

10 ms overlapping between consecutive frames. Then, the following features are extracted: autocorrelation of a filter response, autocorrelation of the envelope of a filter response, cross-channel correlation, and dominant pitch within each time frame.

1. *Correlogram*: For a Time-Frequency unit u_{cm}, its autocorrelation function is given by

$$A_H(c,m,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h(c,mT-n)h(c,mT-n-\tau). \tag{3.9}$$

   were τ is a delay between 0 and 12.5 ms. T is the time shift from one frame to the next and N_c the corresponding number of samples.

2. *Dominant pitch*: Let $s(m,\tau)$ be the summary correlogram at frame m

$$s(m,\tau) = \sum_c A_H(c,m,\tau). \tag{3.10}$$

   We define the dominant pitch period at frame m, $\tau_D(m)$, to be the lag corresponding to the maximum of $s(m,\tau)$ in the plausible pitch range of target speech [2 ms, 12.5 ms].

3. *Envelope Correlogram*: the envelope correlogram is made by computing the autocorrelation of a response envelope

$$A_E)(c,m,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h_E(c,mT-n)h_E(c,mT-n-\tau). \tag{3.11}$$

   Here, $h_E(c,n)$ is the envelope of the output in channel c at time step n. It reveals response periodicities as well as AM rates.

4. *Cross-Channel Correlation*: Wang and Brown [ref] demonstrated that cross correlation between adjacent filter channels indicates whether the filters mainly respond to the same

source or not. For the same T-F unit, the cross channel correlation is calculated as

$$C_H(c,m) = \sum_{\tau=0}^{L-1} \hat{A}_H(c,m,\tau)\hat{A}_H(c+1,m,\tau) \tag{3.12}$$

where $\hat{A}_H(c,m,\tau)$ denotes $A_H(c,m,\tau)$ normalized to zero mean and unity variance, and $L = 201$ corresponds to 12.5 ms, i.e. the maximum delay for A_H. Similarly, the cross-channel correlation of envelopes is calculated as

$$C_E(c,m) = \sum_{\tau=0}^{L-1} \hat{A}_E(c,m,\tau)\hat{A}_E(c+1,m,\tau) \tag{3.13}$$

where $\hat{A}_E(c,m,\tau)$ denotes normalized $A_E(c,m,\tau)$.

These features are used in the following stages.

3.2.1 Initial segregation

In this stage, units are merged into segments based on temporal continuity and cross-channel correlation. Using dominant pitch, these segments are grouped into an initial foreground stream and an initial background stream.

1. *Initial Segmentation*: Only units with some response energy and sufficiently high-cross-channel correlations are considered. A unit u_{cm} is selected if $A_H(c,m,0) > \theta_H^2$ and $C_H(c,m) > \theta_C$, with $\theta_H = 50$ and $\theta_C = 0.985$. Selected neighboring units are iteratively merged into segments. Segments shorter than 30 ms are removed since they unlikely arise from target speech.

2. *Initial Grouping*: This grouping is done through comparing the periodicities of unit responses with dominant pitch. A unit u_{cm} is said to agree with the dominant pitch if

$$\frac{A_H(c,m,\tau_D(m))}{A_H(c,m,\tau_P(c,m))} > \theta_P \tag{3.14}$$

   Here $\theta_P = 0.95$, and $\tau_P(c,m)$ is the delay corresponding to the maximum of $A_H(c,m,\tau)$. is the delay corresponding to the maximum of $A_H(c,m,\tau)$ within the plausible pitch range [2 ms, 12.5 ms]. For any segment, if more than half of its units at a certain frame agree with the dominant pitch, this segment is said to agree with the dominant pitch at this frame. The longest segment is selected as a seed stream. At a certain frame, a segment is said to agree with the longest segment if both segments if both segments agree or both disagree with the dominant pitch. If a segment agrees with the longest segment for more than half of their overlapping frames, its T-F units within the duration of the longest segment is grouped into the seed stream. Otherwise, the segment is grouped into the competing stream. The longest segment is also used to determine which stream correspond to target speech. If it agrees with the dominant pitch for more than half of its frames, it is likely to contain dominant target speech. In this case we refer to the stream containing the longest segment as the foreground stream, S_F^0, and the competing stream as the background stream, S_B^0. Otherwise, the names of the two streams are swapped.

3.2.2 Pitch Tracking and Unit Labeling

To obtain a more accurate pitch contour, the pitch is re-estimated by verifying the pitch contour obtained from S_F^0 witch two psychoacoustically motivated constraints:

1. An accurate pitch period is consistent with the periodicities of individual constituent units. A unit u_{cm} agrees with $\tau_S(m)$ if

$$\frac{A_H(c, m, \tau_S(m))}{A_H(c, m, \tau_P(c, m))} > \theta_P \tag{3.15}$$

   If an estimated pitch period is reliable, at least half of the units in the foreground stream at the corresponding frame must agree with it.

2. The pitch contour of speech changes slowly, so the difference between the reliable pitch periods at frame m and $m+1$ must be less than 20% of themselves. This only is applied to pitch periods satisfying the first constraint.

Once the pitch contour is computed, it is used to label T-F units according to whether target speech dominates the unit responses or not. A unit is labeled by comparing its response periodicity with the estimated pitch period. Then, a unit $u_{c,m}$ is labeled as target speech if

$$\frac{A_H(c, m, \tau_S(m))}{A_H(c, m, \tau_P(c, m))} > \theta_T \tag{3.16}$$

with $\theta_P = 0.85$. This is called the periodicity criterion and works fine with resolved harmonics. Foreground and background streams are subsequently adjusted. A segment is grouped into the foreground stream, now denoted as S_F^2 if it agrees with the new pitch contour of target speech, according to 3.16, for more than half of its length, otherwise, it is put into the background stream S_B^2.

For units responding to multiple harmonics, their responses are amplitude-modulated and as result, the pitch of target speech does not necessarily correspond to the global maximum of the autocorrelation of such a unit in the plausible pitch range. For a filter responding to multiple harmonics of a single harmonic sound source, the response envelope fluctuates at the rate of $F0$ of the source. This is referred as the AM criterion. For each unit, the AM criterion compares the AM rate with the estimated pitch as follows.

1. The response of a gammatone filter is half-wave rectified and band-passed to remove the DC component and all possible harmonics except for the $F0$ component. The rectified and filtered signal is the normalized by its envelope to remove intensity fluctuations of the original mixture. That is

$$\hat{r}(c, n) = \frac{r(c, n)}{r_E(c, n)} \tag{3.17}$$

   where $r(c, n)$ is the rectified and filtered output in channel c at time step n, and $r_E(c, n)$ es the envelope of $r(c, n)$ obtained via the Hilbert transform.

2. The corresponding normalized signal within a T-F unit u_{cm} is model by a single sinusoid with the specified period of $\tau_S(m)$, in order to compare the AM rate with the estimated pitch period. Specifically, setting to 0 the derivative of the square error between $\hat{r}(c, n)$ and this sinusoid
$$\tan \phi_{cm} = -\frac{\sum_{n=0}^{2T-1} \hat{r}(c, mT-n) \sin\left(\frac{2\pi n}{\tau_S(m) f_S}\right)}{\sum_{n=0}^{2T-1} \hat{r}(c, mT-n) \cos\left(\frac{2\pi n}{\tau_S(m) f_S}\right)}. \tag{3.18}$$
where j is the imaginary unit and $f_S = 16$ kHz is the sampling frequency. Note that within $[0, 2\pi)$, there are two solutions for this equation. A unit is labeled as target speech if the following square error is below a certain percentage of the total energy of the corresponding signal
$$\frac{\sum_{n=0}^{2T-1} \left[\hat{r}(c, mT-n) - \cos\left(\frac{2\pi n}{\tau_S(m) f_S} + \phi_{cm}\right)\right]^2}{\sum_{n=0}^{2T-1} \hat{r}^2(c, mT-n)} < \theta_{AM}. \tag{3.19}$$
where θ_{AM} is chosen to be 0.2.

3.2.3 Final Segregation

In this stage, segments corresponding to unresolved harmonics are generated based on temporal continuity and cross-channel envelope correlation. First, T-F units are selected if they are labeled as target speech but do not belong to any segment generated in initial segmentation and their C_E's are greater than 0.985. Selected neighboring units are iteratively merged into segments. Finally, to reduce the influence of noise intrusion, segments shorter than 50 ms are removed. Al the generated segments are added to S_F^2.

The spectra of target speech and intrusion often overlap and some segments generated in initial segmentation contain units where target dominates as well as those where intrusion dominates. A segment in S_F^2 can be further divided into smaller segments so that all the units in a segment have the same label. Then the segments in S_F^2 are adjusted as follows:

- segments with the target label are retained in S_F^2 if they are no shorter than 50 ms.
- segments with the intrusion label are added to S_B^2 if they are no shorter than 50 ms.
- remaining segments are removed from S_F^2 and they become undecided.

Then S_B^2 expends iteratively to include undecided segments in its neighborhood. All the remaining undecided segments are added back to S_F^2.

Finally, individual units that do not belong to either stream are grouped into the foreground stream iteratively if they are labeled as target speech and in the neighborhood of the foreground stream. The result of this is the final segregated stream of target speech, denoted as S_F^3. The remaining units are added to the background stream, yielding S_B^3.

Chapter 4

Stereo SSS Algorithms

In this chapter, two algorithms for stereo music sound source separation are described: the ADRess algorithm [2] and a novel method based on time frequency masking and multilevel thresholding.

4.1 ADRess: Azimuth Discrimination and Resynthesis

Many studio recordings use the panoramic potentiometer in order to achieve image localisation. The pan pot allows a single sound source to be divided into two channels with continuously variable intensity ratios. This makes the source virtually positioned at any point between the two stereo loudspeakers. This localisation is achieved by creating an interaural intensity difference (IID).

The ADRess method applies gain-scaling to one channel so that one source's intensity becomes equal in both left and right channels. A simple subtraction of the channels will cause that source to cancel out out due to phase cancellation. The cancelled source is recovered by first creating a "frequency-azimuth" plane which is then analyzed for local minima along the azimuth axis. It is observed that at some point where an instrument cancels, only the frequencies which it contained will show a local minima. The magnitude and phase of these minima are then estimated and used with an overlap add scheme to resynthesize the cancelled instrument.

4.1.1 Azimuth Discrimination

The mixing process can be expressed as

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \tag{4.1}$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \tag{4.2}$$

where S_j are the J independent sources, Pl_j and Pr_j are the left and right panning coefficients for j^{th} source, and L and R are the resultant left and right mixture channels. It can be seen that the intensity ratio of the j^{th} source, $g(j)$, between the left and right channels can be expressed as

$$g(j) = \frac{Pl_j}{Pr_j} \tag{4.3}$$

Multiplying the right channel, R, by $g(j)$ will make the intensity of the j^{th} source equal in left and right. The operation $L - g(j)R$ will cause the j^{th} source to cancel out. In practice, $L - g(j)R$ is used if the j^{th} source is predominant in the right channel, and $R - g(j)L$ if the j^{th} source is predominant in the left channel. This way, $g(j)$ is always between 0 and 1, and it insures that we are always scaling one channel down in order to match intensities, avoiding distortion caused by large scaling factors.

Recovering of the cancelled source is made in the frequency domain. We divide the stereo mixture into short time frames and carry out an FFT on each:

$$Lf(k) = \sum_{n=0}^{N-1} L(n)e^{-j\frac{2\pi}{N}} \tag{4.4}$$

$$Rf(k) = \sum_{n=0}^{N-1} R(n)e^{-j\frac{2\pi}{N}} \tag{4.5}$$

where Lf and Rf are short time frequency domain representations of the left and right channels respectively. In practice a 4096 point FFT with a Hanning window is used, and a step size of 1024 points. Next, two frequency azimuth planes are created, one for the left channel and one for the right channel. The azimuth resolution, β, refers to how many equally spaced gain scaling values of g will be used to construct the plane. Specifically

$$g(i) = i \times \frac{1}{\beta} \tag{4.6}$$

for all i, where $0 \leq i \leq \beta$, and where i and β are integer values.

Figure 4.1 shows the azimuth domain for a stereo signal of a two partial mixture. Large values of β will lead to more accurate azimuth discrimination but also to a higher computational load. Assuming an N point FFT, the frequency-azimuth plane will be an $N \times \beta$ array for each channel:

$$Az_{R(k,i)} = |Lf(k) - g(i)Rf(k)| \tag{4.7}$$

$$Az_{L(k,i)} = |Rf(k) - g(i)Lf(k)| \tag{4.8}$$

for all i and k where, $0 \leq i \leq \beta$ and $1 \leq k \leq N$.

This process will reveal frequency dependent nulls, corresponding to cancelled sources due to gain scaling. In order to estimate the magnitude of these nulls we redefine equations 4.7 and 4.8 as

$$Az_{R(k,i)} = \begin{cases} Az_{R(k)_{max}} - Az_{R(k)_{min}} & if \quad Az_{R(k,i)} = Az_{R(k)_{min}} \\ 0 & otherwise \end{cases} \tag{4.9}$$

$$Az_{L(k,i)} = \begin{cases} Az_{L(k)_{max}} - Az_{L(k)_{min}} & if \quad Az_{L(k,i)} = Az_{L(k)_{min}} \\ 0 & otherwise \end{cases} \tag{4.10}$$

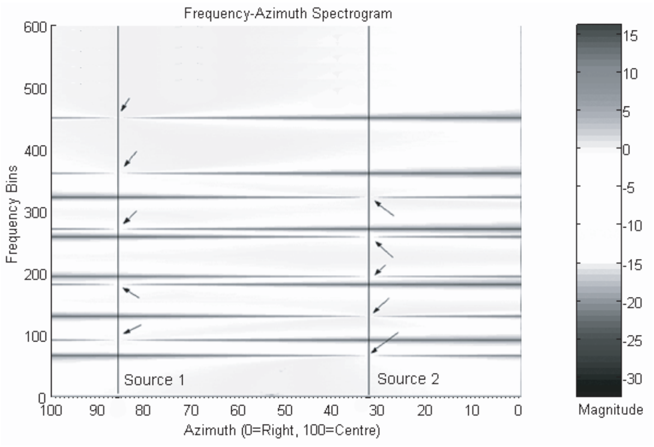


Figure 4.1: The Frequency-Azimuth spectrogram for a mixture of 2 synthetic sources each comprising of 5 non-overlapping partials. The arrows indicate frequency dependent nulls caused by phase cancellation. Extracted from [17].

By doing this, nulls are turned into peaks. However, in practice there is harmonic overlap between the sources to be recovered. Harmony if one of the fundamentals of music, and instruments will be often playing harmonically related notes simultaneously. The result will be a significant harmonic overlap, producing a "frequency-azimuth smearing". This is caused when two or more sources contain energy in a single frequency bin. The apparent frequency dependent null drifts away from a source position and may be at a minimum at a position where there is no source at all. To overcome this problem, an "azimuth subspace width", H, is defined such that $1 \leq H \leq \beta$. This allows us to recover peaks within a given neighborhood. A wide azimuth subspace will result in worse rejection of nearby sources. On the other hand a narrow azimuth subspace will lead to poor resynthesis and missing frequency information.

4.1.2 Resynthesis

In order to resynthesize only one source, a discrimination index d is set to the apparent position of the source. The azimuth subspace width, H, is then set such that the best resynthesis quality is achieved. In practice, the azimuth subspace is centered over the discrimination index such that the subspace spans from $d - H/2$ to $d + H/2$. The peaks for resynthesis are extracted using

$$Y_{R(k)} = \sum_{i=d-H/2}^{d+H/2} Az_{R(k,i)} \quad 1 \leq k \leq N \tag{4.11}$$

$$Y_{L(k)} = \sum_{i=d-H/2}^{d+H/2} Az_{L(k,i)} \quad 1 \leq k \leq N \tag{4.12}$$

The resultant Y_R and Y_L are $1 \times N$ arrays containing only the bin magnitudes pertaining to a particular azimuth subspace as defined by d and H. More specifically, Y_R and Y_L contain the short time power spectrum of the separated source. It should be noted that, if two sources have the same intensity ratio, i.e. they share the same pan position, both will be present in the extracted subspace. This is particularly true of the "center" position. It is a common practice in audio mix down to place a number of instruments here, usually voice and very often bass guitar and elements of the drum kit too.

Keeping the original bin phases in the resynthesis step produce acceptable results. Then, the azimuth subspace is then resynthesized by using an IFFT for each frame. The resynthesized time frames are combined using a standard overlap-add scheme.

4.2 Separation based on Time Frequency Masking and Multilevel Thresholding

We present a method for automatic separation of sources in stereo recordings. As other methods used in stereo source separation, the framework used in this paper is also based on the analysis of the interaural intensity difference (IID) existent between the two observation channels in the STFT domain [8][26]. A basic assumption made by these algorithms is that in the time-frequency transform domain, signal components corresponding to different sources do not overlap significantly [**?**]. This is often called the W-disjoint orthogonality assumption. Whereas some separation methods need specific information about the panning configuration [1] or human attendance [21] for completing the separation process, the method described in this work performs an automatic estimation of the optimum time-frequency masks for different sources. A $\log^{-1}$ weighted histogram and the multilevel extension of the Otsu's thresholding algorithm [14] are used for this purpose.

4.2.1 Stereo Mixing Model

Studio recordings can be modelled as a sum of J amplitude panned sources $s_j(t)$, $j = 1 \ldots J$ convolved with reverberation impulse responses $r_i(t)$ for each channel $i = 1, 2$. The stereo mixture channels can be written as

$$x_i(t) = \left[\sum_{j=1}^{J} a_{ij} s_j(t) \right] * r_i(t) \quad i = 1, 2 \tag{4.13}$$

where a_{ij} are the amplitude panning coefficients used in the stereo mixdown. Assuming a short reverberation impulse response in each channel, the mixture becomes instantaneous:

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t) \quad i = 1, 2. \tag{4.14}$$

To formalize, we denote the STFT's of the channel signals $x_i(t)$ as $X_i(k, m)$, where k is the frequency index and m is the time index. Given the linearity of the STFT, we can write the STFT of each channel as

$$X_i(k,m) = \sum_{j=1}^{J} a_{ij} S_j(k,m) = \sum_{j=1}^{J} S_{ij}(k,m) \quad i = 1,2 \tag{4.15}$$

where $S_{ij}(k,m) = a_{ij}S_j(k,m)$ is the image of source j in channel i in the STFT transform domain.

A source j is said to be panned to the left if $a_{1j} > a_{2j}$. If $a_{1j} < a_{2j}$ the source is said to be panned to the right. If $a_{1j} = a_{2j}$ we say that the source is panned to the center. The mixing model can be also written as

$$x_i(t) = \sum_{p=1}^{J_1} a_{ip}s_p(t) + \sum_{q=1}^{J_2} a_{iq}s_q(t) + \sum_{c=1}^{J_c} a_{ic}s_c(t) \tag{4.16}$$

where J_1 is the number of sources panned to the left, J_2 the number of sources panned to the right and J_c the number of sources panned to the center.

4.2.2 Separation Framework

Most of the stereo separation methods consist in estimating the coefficients used in the mixing process in order to make a clustering of time frequency points that have a similar mixing ratio [8][1][21][26]. The general approach is to define a two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures. Peaks corresponding to the relative attenuation and delay mixing parameters of each source are observed and time-frequency masks are formed for each peak, allowing the separation. This approach has shown to provide good results when dealing with two channel speech mixtures, but insufficient when music recordings are considered. When multiple instruments and singing voice are present, the overlap is much more significant and the mixing ratio varies so much that no clear peaks can be observed in the histogram.

In this work we describe a method based on the previous described framework. We propose to use a perceptual weighted histogram made up with time-frequency points only in the medium frequency range, where the sources are supposed to concentrate their energy. Then, the Fast Multilevel Otsu Algorithm [12] is used for searching the optimal thresholds that maximize the between-class variance of the mixing ratio values.

Mixing Ratio Deviation in Quasi W-Disjoint Orthogonal Sources

If the mixing coefficients are time invariant, the amplitude ratio between the left and right channels for a single source remains constant:

$$\frac{s_{1j}(t)}{s_{2j}(t)} = \frac{a_{1j}}{a_{2j}} \tag{4.17}$$

Usually the W-disjoint orthogonality assumption is made [8]. The sources are said to be W-disjoint orthogonal if they do not overlap in the STFT transform domain. This can be expressed mathematically as

$$S_i)(k,m)S_j(k,m) = 0 \quad \forall i \neq j, \forall k, m \tag{4.18}$$

Thus, only an active source will be present in each time-frequency point, and the ratio between the magnitude of the STFT of the mixture channels will correspond to the ratio between the mixing coefficients of the active source, given by

$$\rho_W(k,m) = \frac{|X_1(k,m)|}{|X_2(k,m)|} = \frac{|S_{1a}(k,m)|}{|S_{2a}(k,m)|} = \frac{a_{1a}}{a_{2a}} \tag{4.19}$$

where the subindex W refers to the W-disjoint orthogonality assumption and a is the index of the active source in the time-frequency point (k,m).

The mixing ratio would uniquely identify the time-frequency components of the sources in the stereo mix only when they are all panned to different locations and do not overlap significantly in the transform domain, as discussed in [1]. In practice, the sources present in the audio signal (and especially in music recordings) are overlapped in time and frequency. This means that there will be a set C of interfering sources which have energy in a shared time-frequency point with a main source of interest s_j:

$$\rho(k,m) = \frac{|X_1(k,m)|}{|X_2(k,m)|} = \frac{|S_{1j}(k,m)| + \sum\limits_{i \in C} |S_{1i}(k,m)|}{|S_{2j}(k,m)| + \sum\limits_{i \in C} |S_{2i}(k,m)|} \tag{4.20}$$

The estimated mixing ratio for the source of interest will correspond to the mixing ratio under the W-disjoint orthogonality assumption plus a deviation produced by the interfering sources:

$$\begin{aligned} \rho(k,m) &= \frac{|S_{1j}(k,m)|}{|S_{2j}(k,m)|} + \frac{\sum\limits_{i \in C} |S_{1i}(k,m)| - \frac{|S_{1j}(k,m)|}{|S_{2j}(k,m)|} \sum\limits_{i \in C} |S_{2i}(k,m)|}{|S_{2j}(k,m)| + \sum\limits_{i \in C} |S_{2i}(k,m)|} \\ &= \rho_W(k,m) + \Delta \end{aligned} \tag{4.21}$$

Taking the logarithm of $\rho(k,m)$, we get

$$\begin{aligned} P(k,m) &= 20 \log\left(\rho(k,m)\right) = 20 \log\left(\rho_W(k,m)\right) \\ &+ 20 \log\left(1 + 10^{(\log \Delta - \log(\rho_W(k,m)))}\right). \end{aligned} \tag{4.22}$$

We call $P(k,m)$ the pan map and it represents the log-mixing ratio of each time-frequency point in the STFT transform domain.

Pan Map Splitting

The first step for the separation of the sources consists in splitting the pan map $P(k,m)$ into two parts corresponding to sources panned to the left and sources panned to the right. This is made by creating two binary masks, one for positive values of the pan map and another one for the negative values

$$U^{(1)}(k,m) = \begin{cases} 1 & if \quad P(k,m) \geq 0 \\ 0 & if \quad P(k,m) < 0 \end{cases} \tag{4.23}$$

$$U^{(2)}(k,m) = \begin{cases} 1 & if \quad P(k,m) \leq 0 \\ 0 & if \quad P(k,m) > 0 \end{cases} \tag{4.24}$$

Multiplying the pan map by these masks, we split $P(k,m)$ into two parts

$$P(k,m) = \sum_{i=1}^{2} P(k,m)U^{(i)}(k,m) = \sum_{i=1}^{2} P^{(i)}(k,m) \tag{4.25}$$

Figure 4.2 shows the mixture spectrograms $X_i(k,m)$, where four sources (piano, guitar, drums and singing voice) are overlapped in the time-frequency transform domain. The original stereo is a instantaneous mixture of these sources sampled at 44.1 kHz. The STFT was carried out using a Hanning window of length 180 ms with 75% overlap. The pan map and the two binary masks obtained for these music mixtures are represented in Figure 4.3.

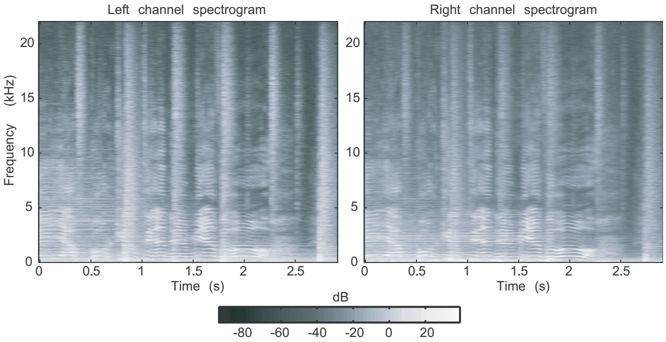


Figure 4.2: Left channel and right channel amplitude spectrograms.

Histogram Formation

The second step consists in estimating the mixing ratios of the sources panned to the left and those panned to the right separately by analyzing the absolute value of the previously calculated pan maps $|P^{(i)}(k,m)|$.

First, $|P^{(1)}(k,m)|$ and $|P^{(2)}(k,m)|$ are normalized, obtaining $Pn^{(1)}(k,m)$ and $Pn^{(2)}(k,m)$. Then, two histograms of L uniform containers in the range $[0,1]$ are formed for $Pn^{(i)}(k,m)$, just taking into account only points in a medium frequency range $[k_{min}, k_{max}]$. The center of each container is given by

$$z_n = \frac{1}{2L}(2n+1) \quad n = 0 \ldots L-1 \tag{4.26}$$

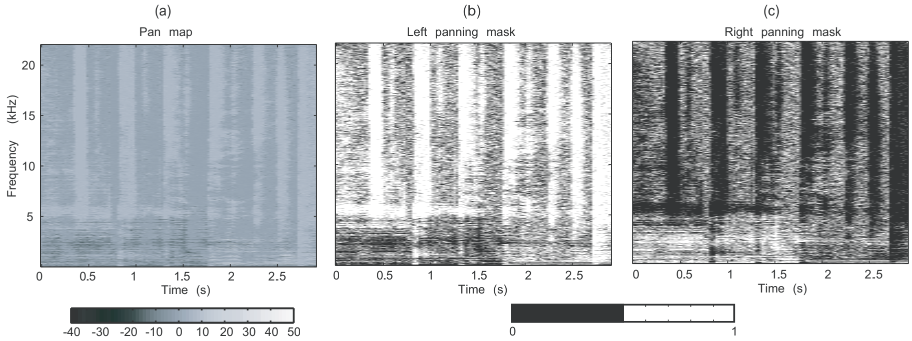


Figure 4.3: (a) Pan map obtained from the spectrograms showed in Figure 4.2. (b) Binary mask for sources panned to the left. (c) Binary mask for sources panned to the right.

We take k_{max} the index of the closest frequency to 4 kHz and k_{min} the index of the closest frequency to 100 Hz. This histogram is calculated as a $\log^{-1}$ weighted sum of the number of points that lie in each of the L containers previously defined. This procedure gives a greater value to points in the lower part of the frequency range of interest:

$$H(n) = \sum_{i \in n} g(k_i), \tag{4.27}$$

where $g(k_i)$ is the weighting factor for a point (k_i, m_i) with value $Pn^{(1)}$ or $Pn^{(2)}$ (depending on the channel considered) in the value range defined by container n. This is a first approximation to perceptual weighting and can be calculated as

$$g(k) = \frac{\log(100)}{\log(100 + k - k_{min})}. \tag{4.28}$$

In the original mixdown of the described example, guitar was panned to the right, drums and piano were panned to the left, and singing voice was positioned at the centre of the azimuth plane. If the sources were perfectly W-disjoint orthogonal, clear peaks corresponding to each source should be easily identified in the histogram containers corresponding to the different mixing ratios of the present sources. In [1] a range of values where the mixing ratios for each source may vary are selected using a Gaussian window. However, the central point of the window must be specified for carrying out the separation. In [21] a human-assisted criterion is used. We propose to use a multilevel thresholding algorithm for selecting the range of values in the histograms corresponding to each source. This way, the sources are extracted automatically by maximizing their inter-class variance defined by Otsu.

Multilevel Thresholding

Thresholding is an important technique for image segmentation which is used for identification and extraction of targets from its background on the basis of the distribution of pixel intensities in image objects. In our separation framework, image segmentation and source extraction from a mixture pan map can be observed from the same point of view. In the separation context, we try to find the thresholds that maximize the inter-class variance of the distribution of mixing ratios over the time-frequency transform domain.

We briefly describe the Otsu's algorithm [14]:

The probability of the $\log^{-1}$ weighted mixing ratio in the middle of container n of the histogram is given by:

$$p_n = \frac{H(n)}{N} \tag{4.29}$$

where $N = \sum_{n=1}^{L} H(n)$.

In the case of bi-level thresholding, the time-frequency points are divided into two classes, c_1 with mixing ratios in the range given by the histogram bins $n \in [1, \ldots, t]$ and c_2 with values within the bins $n \in [t+1, \ldots, L]$. Then, the probability distributions for the two classes are:

$$c_1 : p_1/\omega_1(t), \ldots, p_t/\omega_1(t) \tag{4.30}$$

$$c_2 : p_{t+1}/\omega_2(t), p_{t+1}/\omega_2(t), \ldots, p_L/\omega_2(t) \tag{4.31}$$

where $\omega_1(t) = \sum_{n=1}^{t} p_n$ and $\omega_2(t) = \sum_{n=t+1}^{L} p_n$.

The means for classes c_1 and c_2 are

$$\mu_1 = \sum_{n=1}^{t} n \frac{p_n}{\omega_1(t)} \tag{4.32}$$

$$\mu_2 = \sum_{n=t+1}^{L} n \frac{p_n}{\omega_2(t)} \tag{4.33}$$

Let μ_T be the mean mixing ratio for the whole image. Then:

$$\omega_1\mu_1 + \omega_2\mu_2 = \mu_T \tag{4.34}$$

$$\omega_1 + \omega_2 = 1 \tag{4.35}$$

Otsu defined the between-class variance as:

$$\sigma_B^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2 \tag{4.36}$$

For bi-level thresholding, Otsu verified that the optimal threshold t^* is chosen so that the between-class variance σ_B^2 is maximized, that is:

$$t^* = \text{Arg max}\left\{\sigma_B^2(t)\right\} \quad 1 \leq t \leq L \tag{4.37}$$

The previous formula can be easily extended to multilevel thresholding. Assuming that there are $M-1$ thresholds, $\{t_1, t_2, \ldots, t_{M-1}\}$, which divide the original pan map into M classes: c_1 for $[1, \ldots, t_1]$, c_2 for $[t_1+1, \ldots, t_2]$, ..., c_i for $[t_{i-1}+1, \ldots, t_i]$ and c_M for $[t_{M+1}+1, \ldots, L]$, the optimal thresholds $t_1^*, t_2^*, \ldots, t_{M-1}^*$ are chosen by maximizing σ_B^2 as follows:

$$\{t_1^*, t_2^*, \ldots, t_{M-1}^*\} = \text{Arg max}\left\{\sigma_B^2(t_1, t_2, \ldots, t_{M-1})\right\} \quad 1 \leq t \leq L \tag{4.38}$$

where

$$\sigma_B^2 = \sum_{k=1}^{M} \omega_k(\mu_k - \mu_T)^2, \tag{4.39}$$

with

$$\omega_k = \sum_{n \in c_k} p_n \tag{4.40}$$

$$\mu_k = \sum_{n \in c_k} n \frac{p_n}{\omega(k)} \tag{4.41}$$

The ω_k in Eq. 4.40 is regarded as the zeroth-order cumulative moment of the kth class c_k, and the numerator in Eq. 4.41 is regarded as the first-order cumulative moment of the kth class c_k, that is

$$\mu(k) = \sum_{n \in c_k} np_n. \tag{4.42}$$

Regardless of the number of classes being considered during the thresholding process, the sum of the cumulative probability functions of M classes equals one, and the mean of the mixing ratios considered is equal to the sum of the means of M classes weighted by their cumulative probabilities. The between-class variance in Eq. 4.39 can thus be rewritten as

$$\sigma_B^2 = \sum_{k=1}^{M} \omega_k \mu_k^2 - \mu_T^2. \tag{4.43}$$

Because the second term in Eq. 4.43 is independent of the choice of the thresholds, the optimal thresholds can be chosen by maximizing $\sigma_B'^2$, which is defined as the summation term on the right-hand side of Eq. 4.43:

$$\sigma_B'^2 = \sum_{k=1}^{M} \omega_k \mu_k^2 \tag{4.44}$$

A faster algorithm can be achieved by recursive calculation of Eq. 4.44. [12]. Let us define the look-up tables for the $u - v$ interval:

$$P(u, v) = \sum_{n=u}^{v} p_n \tag{4.45}$$

$$S(u, v) = \sum_{n=u}^{v} np_n \tag{4.46}$$

For index $u = 1$, equations 4.45 and 4.46 can be rewritten as

$$P(1, v + 1) = P(1, v) + p_{v+1} \quad \text{and} \quad P(1, 0) = 0 \tag{4.47}$$

$$S(1, v + 1) = S(1, v) + (v + 1)p_{v+1} \quad \text{and} \quad S(1, 0) = 0 \tag{4.48}$$

From equations 4.47 and 4.48, it follows that

$$P(u, v) = P(1, v) + P(1, u - 1) \tag{4.49}$$

and

$$S(u, v) = S(1, v) + S(1, u - 1) \tag{4.50}$$

Now, the modified between-class variance $\sigma_B'^2$ can be rewritten as

$$\sigma_B^2 = G(1, t_1) + G(t_1 + 1, t_2) + \ldots + G(t_{M-1} + 1, L), \tag{4.51}$$

where the modified between-class variance of class c_i is defined as

$$G(t_{i-1} + 1, t_i) = \frac{S(t_{i-1} + 1, t_i)^2}{P(t_{i-1} + 1, t_i)}. \tag{4.52}$$

The search range for the maximal $\sigma_B'^2$ is $1 \leq t_1 \leq L - M + 1, t_1 + 1 \leq t_2 \leq L - M + 1, \ldots, t_{M-1} + 1 \leq t_{M-1} \leq L - 1$.

The final thresholding values will be those in the middle of containers $n = t_i^*$:

$$Th_i = z_n|_{n=t_i^*} \tag{4.53}$$

Binary Masking

Once the optimum thresholds have been calculated for $Pn^{(i)}$, we are able to define the binary masks corresponding to each class.

Let's call $Th_i^{(1)}$ and $Th_i^{(2)}$ the optimum thresholds for sources panned to the left and right channels, respectively. Figure 4.4 shows the optimum thresholds found for a case where three different classes are considered in each channel histogram. Note that we can search for an arbitrary number of classes in each channel, even if the number of sources panned to that channel is lower. When this happens, a refinement step for clustering several classes to a same source must be carried out. This will be further discussed in the next step.

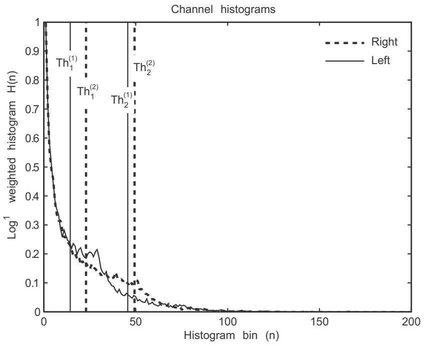


Figure 4.4: Optimum thresholds.

The binary masks for sources panned to the left, are given by

$$U_i^{(1)}(k, m) = \begin{cases} U^{(1)}(k, m) & \text{if} \quad Th_{i-1}^{(1)} < Pn^{(1)}(k, m) \leq Th_i^{(1)} \\ 0 & \text{elsewhere} \end{cases} \tag{4.54}$$

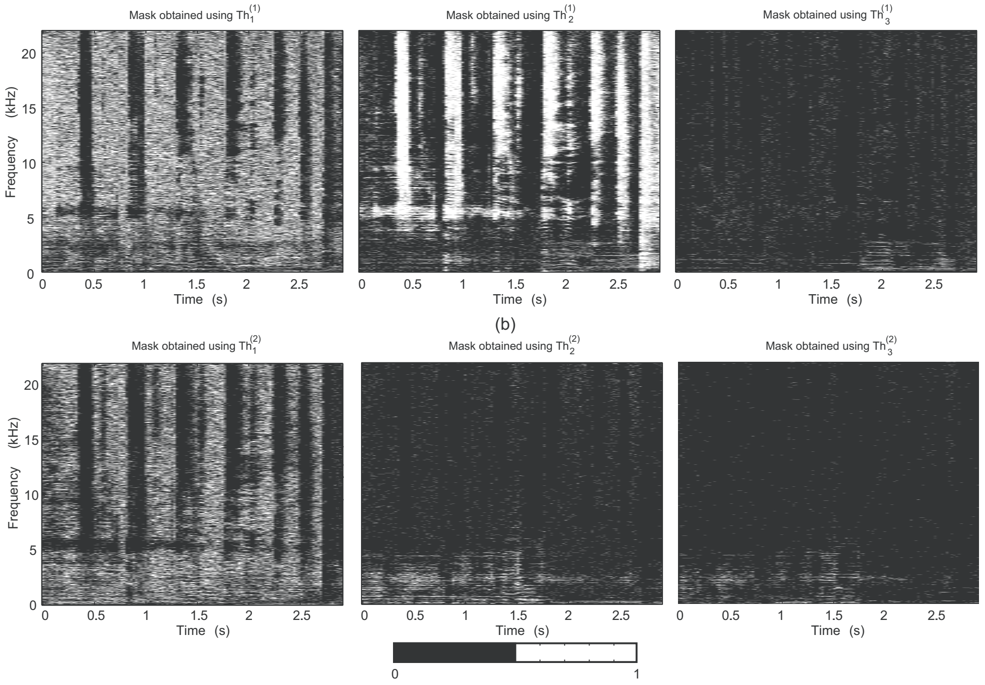
with $i = 1 \ldots M_1$, being M_1 the number of classes to be estimated in the left channel histogram, $Th_0^{(1)} = 0$ and $Th_{M_1}^{(1)} = 1$.

Similarly, for the right channel:

$$U_i^{(2)}(k,m) = \begin{cases} U^{(2)}(k,m) & \text{if} \quad Th_{i-1}^{(2)} < Pn^{(2)}(k,m) \leq Th_i^{(2)} \\ 0 & \text{elsewhere} \end{cases} \tag{4.55}$$

with $i = 1 \ldots M_2$, being M_2 the number of classes to be estimated in the right channel histogram, $Th_0^{(2)} = 0$ and $Th_{M_2}^{(2)} = 1$.

Figure 4.5 (a) and (b) shows the masks formed by applying the obtained thresholds to $Pn^{(1)}$ and $Pn^{(2)}$, respectively.



(b)

Figure 4.5: Primary binary masks. (a) Masks obtained by applying the thresholds in the first histogram to the left binary mask. (b) Masks obtained by applying the thresholds in the second histogram to the right binary mask.

Class Reassignment

As already stated, there's no restriction in the multilevel thresholding process for defining the number of classes M_i in $Pn^{(i)}$. This means that, if the number of sources panned to the left (or right) is lower than the number of classes defined in the thresholding step ($M_i > J_i$), then more than one mask may correspond to the same source. Independently of the number of classes considered, when a source is panned to the center, there will be always two masks corresponding to that source: $U_1^{(1)}$ and $U_1^{(2)}$. In fact, two masks corresponding to a same source are always azimuth adjacent, which simplifies the reassignment step.

First, the masks are azimuth ordered to form a single set of masks:

$$\mathbf{B} = \{B_i, B_2, \ldots, B_{M_1+M_2}\} = \left\{U_{M_1}^{(1)}, \ldots, U_1^{(1)}, U_1^{(2)}, \ldots, U_{M_2}^{(2)}\right\}. \tag{4.56}$$

Although many ways for comparing binary images can be used, we propose a simple way for doing it just by taking a $N \times M$ grid for each mask B_i and computing the number of non-zero points m_n in each cell. This way, a vector for each mask $\mathbf{m}_i = [m_1 \ m_2 \ldots m_{N\times M}]^T$ is formed. Then, we calculate the mean distance between all the adjacent vectors $\mathbf{m}_i$ and $\mathbf{m}_{i+1}$:

$$d_{i,i+1} = \frac{1}{N \times M} \sum_{n=1}^{N\times M} |\mathbf{m}_i(n) - \mathbf{m}_{i+1}(n)| \quad i = 1, \ldots, M_1 + M_2 - 1 \tag{4.57}$$

If $d_{i,i+1}$ is a local or absolute minimum of the whole distances sequence, then their corresponding masks are added: $B_i' = B_i \cup B_{i+1}$. After this reassignment step, a set of $J' \leq M_1 + M_2$ different masks are available for retrieving the original sources. In Figure 4.6, the reassigned masks ordered in azimuth (from left to right) are shown.

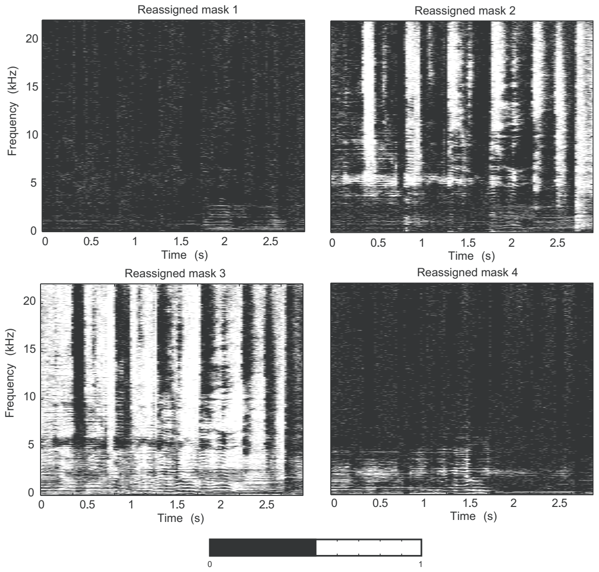


Figure 4.6: Masks obtained after the reassignment step.

Source image retrieval

We can estimate the source images in each channel just applying the calculated masks to the STFT of each channel, conserving the phase information of the mixture:

$$\hat{S}_{ij}(k,m) = |X_i(k,m)| B_j' \, e^{j\angle X_i} \quad j = 1 \ldots J', \ i = 1, 2 \tag{4.58}$$

The estimated sources in time domain will be

$$\hat{s}_j = \mathrm{STFT}^{-1}\left\{S_{1j} + S_{2j}\right\} \quad j = 1 \ldots J'. \tag{4.59}$$

Refinement Step

A refinement step can be carried out for reassigning inter-source residuals in the separated sources by applying the described method recursively. For a separated source $\hat{S}_{ij}(k,m)$, we can calculate its normalized pan map as

$$Pn_j^{(1)} = Pn^{(1)}B'_j \quad \text{or} \quad Pn_j^{(2)} = Pn^{(2)}B'_j \tag{4.60}$$

depending on if the source is panned to the left or to the right. A histogram for Pn_j is carried out as explained in Subsection 4.2.2. Next, a bilevel thresholding ($M = 2$) is applied to the pan map, segregating the initial mask B'_j into two masks, as in Subsections 4.2.2 and 4.2.2. At this time, we may have two masks, one of them corresponding to the primary source in $\hat{S}_{ij}(k,m)$, and another one corresponding to a residual of one of the separated sources adjacent to the one considered: $\hat{S}_{i(j-1)}(k,m)$ or $\hat{S}_{i(j+1)}(k,m)$.

After this reassigning step, the final estimated sources are recovered by applying this masks to the STFT of the mixture channels and calculating the inverse STFT of the result. We show the obtained waveforms in Figure 4.7 (a), and the original waveforms of the sources in Figure 4.7 (b). The similarity between the separated sources and the original is obvious.

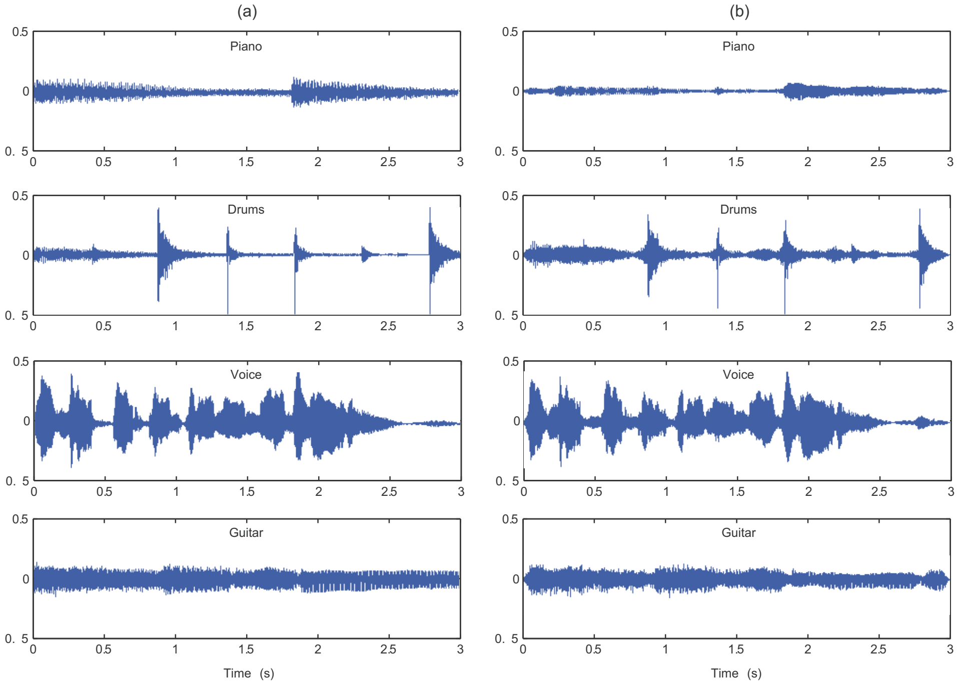


Figure 4.7: Waveform results. (a) Separated sources. (b) Original sources.

Practical Considerations

Although the above separation framework can be extended to an arbitrary number of sources, a practical limit is always present when applying the described processing to a stereo music mixture. This is again a consequence of the mixing process and the W-disjoint orthogonality assumption, which is far from being true for audio sources (for speech mixtures is a more realistic assumption). Note that if two different sources are mixed with the same pan, then they will be extracted as a unique source, as they both will have the same mixing ratio. Moreover, if many sources are present although panned to different azimuth positions, as each time-frequency point is assigned to a different source, the recovered sources will be plenty of artifacts due to the non linear filtering process. This makes not very useful to search for more than three classes in each normalized pan map ($M_1 = M_2 = 3$).

Performance Evaluation

Some performance measures in source separation processing have already been described in the literature. In [17] the objective performance evaluation of sound source separation algorithms is well discussed. In some applications it may be relevant to allow more or less distortions, not necessarily related to the theoretical indeterminacies of the problem. The evaluation procedure developed in [17] takes into account the application for which a given separation algorithm is oriented. For example, in musical applications, it may be important to recover the sources up to a simple gain since arbitrary filtering modifies the timbre of musical instruments. The assumptions made for applying the performance criteria are:

- the true source signals are known,
- a family of allowed distortions is chosen

The computation of the criteria involves two successive steps. First, $\hat{s}_j$ is decomposed as

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif}, \tag{4.61}$$

where $s_{target} = f(s_j)$ is a version of s_j modified by an allowed distortion $f \in \mathcal{F}$, and where e_{interf}, e_{noise} and e_{artif} are respectively the interferences, noise and artifacts error terms. From this decomposition, some numerical performance criteria are defined. The Source to Distortion Ratio

$$\text{SDR} := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}, \tag{4.62}$$

the Source to Interferences Ratio

$$\text{SIR} := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2}, \tag{4.63}$$

the Sources to Noise Ratio

$$\text{SNR} := 10 \log_{10} \frac{\|s_{target} + e_{interf}\|^2}{\|e_{noise}\|^2}, \tag{4.64}$$

Table 4.1: Objective measures with time invariant gain distortion allowed

	SDR		SIR		SAR	
$\hat{s}_j$	ADRess	MLTS	ADress	MLTS	ADRess	MLTS
Piano	-0.2	-2.7	28.4	15.7	-0.2	-2.5
Drums	5.1	3.8	19.6	11.2	5.3	5.0
Voice	-2.3	10.4	13.1	20.6	-2.0	10.9
Guitar	0.1	4.2	13.0	16.9	0.6	4.5

Table 4.2: Objective measures with time invariant filtering distortion allowed

	SDR		SIR		SAR	
$\hat{s}_j$	ADRess	MLTS	ADress	MLTS	ADRess	MLTS
Piano	-0.2	-2.1	19.8	10.9	0.3	-1.5
Drums	5.4	3.9	18.5	10.3	5.7	5.4
Voice	-1.9	10.5	11.1	19.4	-1.4	11.2
Guitar	0.4	4.8	11.9	16.9	1.0	5.2

and the Sources to Artifacts Ratio

$$\mathrm{SAR} := 10 \log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2}, \tag{4.65}$$

These measures can be interesting for comparing several algorithms. Given a family of allowed distortions, the SIR and SAR are valid as performance measures regarding two separate goals: rejection of the interferences and absence of forbidden distortions or "burbling" artifacts. The SNR is valid as a measure of rejection of the sensor noise. The SDR can be seen as a global performance measure.

In this work, we have compared the ADRess algorithm for music separation [2] with the described multilevel thresholding separation method (MLTS), using the same window length and overlap values (180 ms and 0.75%). For that purpose, we have used the MATLAB toolbox *BSS_EVAL* [6], distributed online under the GNU Public License. Table 4.1 shows the SDR, SIR and SAR for the estimated sources using both algorithms when only a time invariant gain distortion is allowed. Table 4.2 shows the evaluated performance when a time invariant filtering is considered (128 taps allowed in the distortion filter). SNR was not considered because no noise was assumed. As we can see in the tables, similar results are obtained allowing both distortions.

The results show how the source panned to the center (voice) is the one extracted with the higher SIR and SAR, as it was expected from the study of the deviation error in [1]. This source and the source panned to the right (guitar) are better extracted using the MLTS method than the ADRess algorithm, which presents periodic gain artifacts. These artifacts are debt to the fact that no cancellations are found in time frames where the source has little energy, producing a noise gate effect. In the MLTS method, the residuals in the extracted sources make the listening more comfortable and this would probably affect positively in subjective evaluation tests. The evaluated tracks can be listened to at http://personales.upv.es/macoser1.

Chapter 5

Sound Scene Resynthesis

This chapter describes how an acoustic scene can be resynthesized by applying SSS algorithms to sound mixtures and feeding the obtained tracks to the WFS rendering algorithm. The resulting perceived quality is evaluated.

5.1 Subjective Evaluation

As we have seen in the previous chapter, some criteria have been proposed in the literature for evaluating separation algorithms from an objective point of view. Although they are related to the perceived audio quality in many cases, they do not model auditory phenomena of loudness weighting and spectral masking.

In [18] some guidelines for subjective evaluation of separation algorithms are given and an adaptation of the MUSHRA standard is proposed. In the WFS framework, subjective evaluation tests should take into account that spatial positioning of the sources is an additional parameter of interest. In order to evaluate the quality of resynthesized scenes with the separated sources, we have compared them with reference scenes created from originally separated signals, where spatial configuration of the sources is kept the same. For this work, we have resynthesized several acoustic scenes by separation of monaural and stereo recordings. Three main case studies for different acoustic scenes have been proposed, in order to apply specific algorithms to specific mixtures. In the first two scenes, both monaural and stereo mixtures have been employed in the experiments. In the third one, only a monaural mixture has been tested.

- Scene 1 is a mixture made up of three ambient sounds: an ambulance (left), a car horn (right) and water dripping sound (front).
- Scene 2 is a music recording of a pop song made up of three sources: singing voice (front), piano (right) and drums (left).
- Scene 3 consists of the same ambulance and car horn of scene 2 (left) mixed up with male speech (right)(monaural).

The 96 loudspeaker array used in the experiments can be seen in Figure 5.1

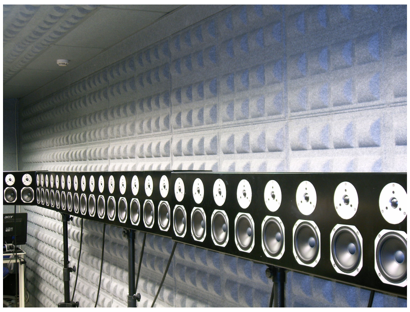


Figure 5.1: WFS array in the GTAC research group.

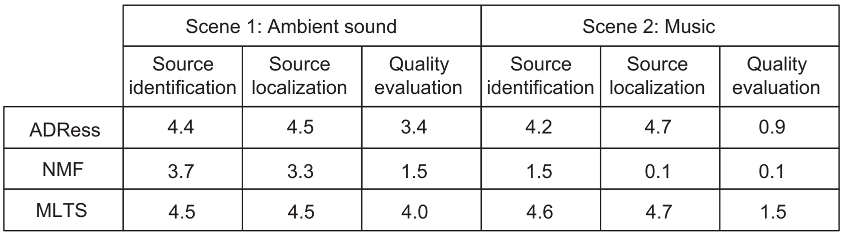
Subjective evaluation was carried out by means of listening tests. The test was performed employing a jury composed of 20 people. The subjects sit in front of the WFS array and the different scenes are presented to them successively. First, an acoustic scene is presented using separated sources obtained from one of the implemented algorithms previously described. Then, they are asked several questions related to the number of sources they can notice and also to the perception of their spatial location. After this, the scene is presented again with the original sources and they are asked the same questions. The last step of the test consists of evaluating the quality of the resynthesized scene by listening to it again once the original scene has been presented. Specifically, three aspects are considered in relation to the test procedure previously commented

1. Source identification: ability of identifying the number of sources present in the mixture.
2. Source localization: ability of identifying the direction of arrival of the sources.
3. Quality evaluation: subjective sound quality of the resynthesized scene in comparison to the reference one.

The allowed score for the third test is: excellent (5), good (3,75), fair (2,5), poor (1,25) and bad (0). For the first and second tests, subjects answers are given a score by comparing their answers in the scene composed using sources separated by means of the algorithm under test with their answers in the reference scene. The maximum score is always given when the answer in both cases (with separated sources and original sources) is the same. The final score for source identification and localization is the mean of the whole scores of the jury.

5.2 Results

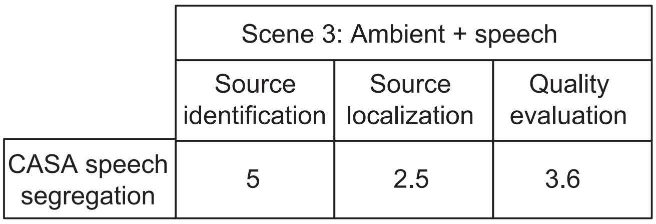
Results of the tests are given in Table 5.2. It shows the score for the ADRess, MLTS and NMF algorithms in case of ambient sound and music. Scene 3 was only processed with the CASA algorithm. The NMF algorithm works worse than the others algorithms, both in the ambient scene and in the music scene. It must be taken into account that stereo algorithms take profit of two observation signals and the NMF algorithm works only with a mono signal. Moreover, it is interesting to appreciate the difference in subjective quality evaluated by the subjects between music and ambient sound. With the ADRess algorithm, subjective quality for Scene 1 (ambient sound) was 3.4, but 0.9 for Scene 2 (music). Similar results are obtained with the MLTS method. This difference shows how subjects tend to be more critical in their evaluation when music is being played, especially when singing voice is present. Source identification and localization is quite good for all the algorithms algorithms in the ambient sound scene but poorer when the NMF algorithm is applied to music.



	Scene 1: Ambient sound			Scene 2: Music		
	Source identification	Source localization	Quality evaluation	Source identification	Source localization	Quality evaluation
ADRess	4.4	4.5	3.4	4.2	4.7	0.9
NMF	3.7	3.3	1.5	1.5	0.1	0.1
MLTS	4.5	4.5	4.0	4.6	4.7	1.5

Figure 5.2: Results for ADRess, MLTS and NMF subjective evaluation in the WFS system

The CASA speech segregation algorithm was applied to Scene 3 in order to segregate speech from the other sounds in the mixture. Table 5.3 shows the scores obtained from the listening tests. The monaural speech segregation algorithm was evaluated good in terms of quality, but all of the subjects noticed that speech was not completely coming from a unique direction. This is because the algorithm was thought to segregate voiced speech (vowels), leaving as background fricative consonants and high frequency components of speech. These background components disturbed the perception of speech making confusing its spatial location.



	Scene 3: Ambient + speech		
	Source identification	Source localization	Quality evaluation
CASA speech segregation	5	2.5	3.6

Figure 5.3: Results for the CASA speech segregation algorithm.

Chapter 6

Summary and Conclusions

In this work, one of the difficulties of the Wave-Field Synthesis systems to be widely deployed has been addressed. The audio signal corresponding to each source in an acoustic scene must be available for making its resynthesis possible. The difficulty resides in the fact that most of the commercial recorded material is in stereo format and there is no possibility to obtain the original multitrack recording. In this work, we have proposed the use of sound source separation techniques to overcome this problem, studying its limitations and discussing the applicability of different separation methods in 3D sound reproduction systems.

The presented work can be summarized as follows:

- First, a description of the sound source separation problem has been carried out, putting special emphasis on its applications, traditional approaches and current research lines.

- Next, some algorithms for sound source separation (monarual and stereo) have been described, implemented and tested in order to resynthesize acoustic scenes in a WFS system: a Non-negative Matrix Factorization algorithm, a Computational Auditory Scene Resynthesis algorithm for speech segregation, a stereo algorithm based on Azimuth Discrimination and Resynthesis, and, finally, a novel method based on Time-Frequency Masking and Multilevel Thresholding, developed by the author.

- A subjective testing campaign involving a jury of 20 people was carried out. The results show that the perceived subjective quality vary with the nature of the scene, being music more critical than ambient sound.

The algorithms used in this work do not give high quality separated signals for hi-fi applications. Nevertheless, masking effects in the WFS reproduction stage relax the quality needed in the separation if they are spatially mixed again. This rises positive hopes for developing full stereo to 3D reproduction systems making use of current and future signal separation methods and, although the results are not definitive, they open a research line to work further.

Chapter 7

Acknowledgments

This work was supported by the Ministry of Education and Science under the project TEC2006-13883-C04-01 and FEDER funds.

Bibliography

[1] AVENDANO, C. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, October 2003).

[2] BARRY, D., LAWLOR, B., AND COYLE, E. Sound source separation: Azimuth discrimination and resynthesis. In *7th Conference on Digital Audio Effects (DAFTX 04)* (2004).

[3] BERKHOUT, A. J. A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36 (1988), 977–995.

[4] BERKHOUT, A. J., DE VRIES, D., AND VOGEL, P. Acoustic control by wave field synthesis. *Journal of the Acoustic Society of America 93* (1993), 2765–2778.

[5] BREGMAN, A. *Auditory Scene Analysis*. 1990.

[6] FÉVOTTE, C., GRIBONVAL, R., AND VINCENT, E. Bss_eval toolbox user guide. Tech. Rep. 1706, IRISA, Rennes, France, April 2006.

[7] HU, G., AND WANG, D. L. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks 15* (2004), 1135–1150.

[8] JOURJINE, A., RICHARD, S., AND YILMAZ, O. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00* (Turkey, 2000), vol. 5, pp. 2985–2988.

[9] JUTTEN, C., AND BABAIE-ZADEH, M. Source separation: principles, current advances and applications. In *German-French Institute for Automation and Robotic Annual Meeting, IAR* (Nancy, France, November 2006).

[10] LEE, C. K., AND CHILDERS, D. G. Cochannel speech separation. *Journal of the Acoustical Society of America* (1988).

[11] LI, Y., AND WANG, D. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing 15* (2007), 1475–1487.

[12] LIAO, P., CHEN, T., AND CHUNG, P. A fast algorithm for multilevel thresholding. *Journal of Information Science and Engineering 17* (2001), 713–717.

[13] O'GRADY, P., PEARLMUTTER, B., AND RICKARD, S. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology (IJIST)* (2005).

[14] OTSU, N. A threshold selection method from gray-level histogram. *IEEE Transactions on System Man Cybernetics SMC-9*, 1 (1979), 62–66.

[15] QUATIERI, T., AND DANISWEICZ, R. G. An approach to co-channel talker interference supression using a sinusoidal model for speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing 38*, 1 (1990).

[16] TORKKOLA, K. Blind separation for audio signals: are we there yet? In *Workshop on Independent Component Analysis and Blind Signal Separation* (1999).

[17] VINCENT, E., GRIBONVAL, R., AND FÉVOTTE, C. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing 14*, 4 (2006), 1462–1469.

[18] VINCENT, E., JAFARI, M. G., AND PLUMBEY, M. D. Preliminary guidelines for subjective evaluation of audio source separation algorithms. In *ICA Research Network Workshop* (University of Liverpool, September 2006).

[19] VINCENT, E., AND RODET, X. Music transcription with isa and hmm. In *5th International Symposium on Indepedent Component Analysis and Blind Signal Separation* (Granada, Spain, 2004).

[20] VINCENT, E., RODET, X., ROBEL, A., FÉVOTTE, C., LE CARPENTIER, E., GRIBONVAL, R., BENAROYA, L., AND BIMBOT, F. A tentative typology of audio source separation tasks. In *ICA* (2003).

[21] VINYES, M., BONADA, J., AND LOSCOS, A. Demixing commercial music productions via human-assisted time-frequency masking. In *Audio Engineering Society 120th Convention* (Paris, France, May 2006).

[22] VIRTANEN, T. Accurate sinusoidal model analysis and parameter reduction by fusion of components. In *110th Audio Engineering Society Convention* (Amsterdam, Netherlands, 2001).

[23] VIRTANEN, T. *Signal Processing Methods for Music Transcription*. 2006.

[24] VIRTANEN, T. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, November 2006.

[25] WANG, D. L., AND BROWN, G. J. *Computational Auditory Scene Analysis*. Wiley-Interscience, 2006.

[26] YILMAZ, O., AND RICKARD, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing* (2003).

Audio Engineering Society

Convention Paper

Presented at the 122nd Convention
2007 May 5–8 Vienna, Austria

On the application of Sound Source Separation to Wave-field Synthesis

Máximo Cobos[1], Jose J. López[2]

[1] iTeAM, Technical University of Valencia, Valencia, 46022, Spain
macoser1@iteam.upv.es

[2] iTeAM, Technical University of Valencia, Valencia, 46022 Spain
jjlopez@dcom.upv.es

ABSTRACT

Wave-field Synthesis (WFS) is a spatial sound reproduction system that can synthesize an acoustic field in an extended area by means of loudspeaker arrays. Spatial positioning of virtual sources is possible but requires separated signals for each source to be feasible. Despite most of the music is recorded in separated tracks for each instrument, in the stereo mixdown process this information is lost. Unfortunately, most of the existing recorded material is in stereo format. In this paper we propose to use Sound Source Separation techniques to overcome this problem. Existing algorithms are yet far from perfection resulting in audible artifacts that clearly reduce the quality of the resynthesized sources in practice. Despite these artifacts, when separated sources are mixed again by a WFS system they are masked by other sounds. The utility of different separation algorithms and the subjective results are discussed as well.

1. WAVE-FIELD SYNTHESIS

Wave-field Synthesis (WFS) is a spatial sound reproduction system that can synthesize a realistic acoustic field in an extended area by means of loudspeaker arrays combined with advanced digital signal processing techniques. The sweet spot effect typical of other spatial sound systems as 5.1 surround is eliminated, obtaining more suitable listening areas. Figure 1 shows an interpretation of the WFS system [1][2].

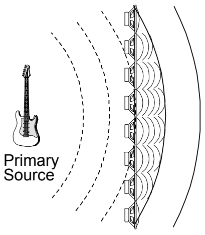


Figure 1 Loudspeaker array (secondary sources) can synthesize the acoustic field created by a primary source

In order to recreate an acoustic scene by means of WFS, the different virtual sources (voices, instruments, etc) that compose the scene are positioned in different space locations as shown in Figure 2. Using the WFS synthesis algorithm the individual excitation signals for each loudspeaker in the array are computed from the individual signals of each instrument in the scene. Therefore, the sound signal of each sound source is needed in the synthesis stage. This presents many advantages because it makes independent the loudspeaker set-up from the sound scene.

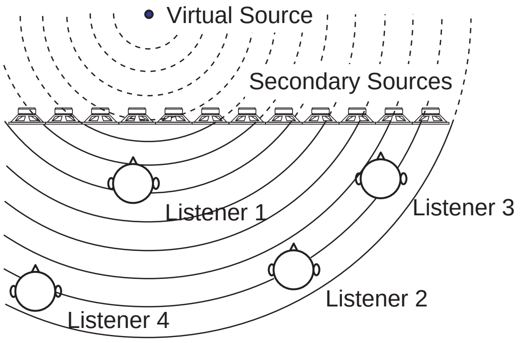


Figure 2 Virtual sources can be located anyplace in the space, making all the listeners perceive it with spatial fidelity

However, the requirement of having the separate sound signal for every source can be a drawback in many cases. Despite most of the music is recorded in the studio in separated tracks for each instrument, in the stereo mixdown process this information is lost. Unfortunately, most of the existing recorded material is only available in stereo format, and there is no possibility to obtain the original multitrack recording.

In this paper we propose to use Sound Source Separation (SSS) techniques to overcome this problem. The scheme in Figure 3 proposes the method for obtaining the signals that feed the WFS rendering algorithm. Different separation algorithms have been subjectively tested with different source materials in the 96 loudspeakers WFS array developed in our research group. The paper is structured as follows: in section 2 an overview of the sound separation algorithms and applications is given, following in section 3 with a review of the different kinds of algorithms to be tested; in section 4 different experiments based on listening tests are carried out.

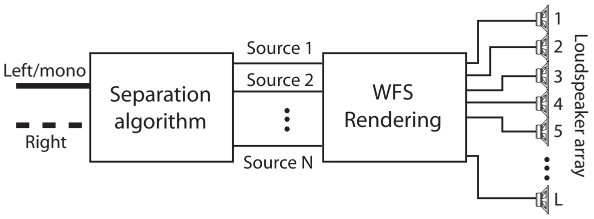


Figure 3 Application of SSS to WFS

2. SOUND SOURCE SEPARATION

2.1. Sound Source Separation Applications

During the last years, Sound Source Separation has been a subject of intense research. It refers to the task of estimating the signals produced by the individual sound sources from a complex acoustic mixture [3][4][5]. Although human listeners are able to perceptually segregate one sound source from an acoustic mixture, “machine listening” systems are still a great challenge.

SSS has a large number of potential applications: high quality separation of musical sources, signal/speech enhancement, multimedia documents indexing, speech recognition in a “cocktail party” environment or source localization for auditory scene analysis. However, current limitations in the existing methods might render some applications if not impossible, at least impractical. A given separation algorithm may perform well on some tasks and poorly on others. It’s because of this fact, that depending on the application, various factors affect the difficulty of the separation, and distinct criteria may be used to evaluate the performance of an algorithm.

Depending on the application, we could be interested in each individual extracted source or maybe just in extracting one source from the mixture (the target source). For example, extraction of singing voice from a song would be an important achievement [6], not just for remixing purposes, but areas like automatic lyrics recognition, singer identification or music information retrieval.

This paper is focused to an Audio Quality Oriented (AQO) application [7]. This means that the extracted sources will be listened to after the separation. In the case of this work, the main purpose will be to examine the possibilities offered by current audio source separation techniques applied to WFS systems. Positioning different sources in different space locations is well accomplished when separated tracks for each source are available. Most of the commercial music productions are recorded this way, but clean information of each source is lost in the mixing process. SSS techniques are the only way to recover the maximum possible information of the different sources.

Although separation algorithms produce resulting signals with plenty of artifacts, they might have less importance when separated sources are mixed again in a WFS system. The isolated tracks for each instrument present artifacts that include mainly, inter-source crosstalk and metallic sound. However, when listening to these tracks all together processed with the WFS system, masking mechanisms are involved. This can make the audition of the resynthesized scene perceptually acceptable even if the separation methods applied are not very sophisticated or flawy.

2.2. Traditional Approaches

The main traditional approaches to the source separation problem have always been beamforming and independent component analysis (ICA). Beamforming achieves sound separation by using the principle of spatial filtering. The aim of beamforming is to boost the signal coming from a specific direction by a suitable configuration of a microphone array at the same time that signals coming from other directions are rejected. The amount of noise attenuation increases as the number of microphones and the array length increase. With a properly configured array, beamforming can achieve high-quality separation.

Independent component analysis models the mixture signal as a standard form of linear superposition of source signals. A mixing model of the form $x(t)=\mathbf{A}s(t)$ is assumed, where $s(t)$ is a vector of unknown source signals, **A** is a mixing matrix, and $x(t)$ is a vector of the mixed signals recorded by several sensors. The main assumption in ICA is that sources involved in the mixing process are statistically independent. The separation problem consists in estimating the unmixing matrix (inverse of **A**). Separation results with ICA are excellent when the assumptions are satisfied, but this not always happen with audio signals [5]. In addition, the number of sensors should be at least equal to the number of sources to be separated. Another fundamental limitation is that the mixing matrix **A** needs to be stationary for a period of time. This assumption is difficult to satisfy in situations in which sound sources slightly move or the environment (acoustic path) changes.

The above techniques are useful just when several observations of the mixture are available. For WFS scene recreation, it would be much more interesting to develop specific algorithms for monaural or stereo recordings. We should concentrate on separation methods where the sources to be separated are not known in advance. These algorithms are based in common properties of real-world sounds, like continuity, sparseness or their harmonic spectral structures.

2.3. One-channel and Stereo Sound Source Separation

The first works on one-channel sound source separation concentrated on the separation of speech signals [8][9]. Analysis and processing of music signals have recently received increasing attention [10][11]. Generally speaking, music is more difficult to be separated than speech. Musical instruments have a wide range of sound production mechanisms, and the resulting signals have a wide range of spectral and temporal characteristics. Even though the acoustic signals are produced independently in each source, it is their consonance and interplay which makes up the music [12]. This results in source signals which depend on each other, which may cause some separation criteria, such as statistical independence to fail.

Approaches used in one-channel sound source separation which do not use source-specific prior knowledge can be roughly divided into three categories, following the classification proposed in [12]:

- **Model based inference**: These methods use a parametric model of the sources to be separated, and the model parameters are estimated from the observed mixture signal. In music applications, the most commonly used parametric model is the sinusoidal model. The model easily enables the prior information of harmonic spectral structure, which makes it the most suitable for the separation of pitched musical instruments and voiced speech [13].

- **Unsupervised learning**: Unsupervised learning methods apply a simple non-parametric model, and use less prior information of the sources to be estimated. Instead, they try to learn the source characteristics from the observed data. The algorithms can apply information-theoretical principles, such as statistical independence between sources. Algorithms which are used to estimate the sources are based on independent subspace analysis [14], non-negative matrix factorization [12], and sparse coding [4].

- **Computational Auditory Stream Analysis (CASA):** CASA methods [15] are based in the ability of humans to perceive and recognize individual sound sources in a mixture referred to as auditory scene analysis [16]. Computational models of this function typically consist of two main stages. First, the mixture signal is decomposed into its elementary time-frequency components. Then, these components are organized and grouped to their respective sound sources. Even though our brain does not resynthesize the acoustic waveforms of each source separately, the human auditory system is a useful reference in the development of one-channel sound source separation systems, since it is the only existing system which can robustly separate sound sources in various circumstances.

Apart from monaural techniques, other approaches have been made to the problem of source separation in music recordings taking advantage of the stereo mixing process. This is the case of the ADRess algorithm [17], which is able to distinguish different sources, analyzing the difference signal from the left and right channels. This is made by searching for minima in planes created from frequency and panning information. Obviously, if one-channel SSS could be achieved, the stereo problem would be solved just working on each channel independently.

3. ALGORITHMS IMPLEMENTED AND TESTED

Before start testing the behavior of SSS algorithms on WFS systems, it is needed to select which one can be interesting or feasible to implement. We have tried to use a representative set of separation algorithms (for monaural and stereo material) from different separation techniques and different case studies. Specifically, for monaural recordings we have tried a Non-negative Matrix Factorization algorithm (NMF) with temporal continuity and sparseness criteria [12] and a Speech Segregation Algorithm based in CASA [18]. For stereo recordings, we have used the ADRess algorithm [17].

3.1. Non-negative Matrix Factorization

The NMF algorithm is based on minimizing the reconstruction error between the magnitude spectrogram of the observed signal and a signal model. The signal model is $\mathbf{X} \approx \mathbf{BG}$, being $\mathbf{X}$ the spectrogram of the mixture, $\mathbf{B}$ the basis function matrix and $\mathbf{G}$ the temporal gain matrix of each basis function. $\mathbf{B}$ and $\mathbf{G}$ are restricted to be entry-wise non-negative. Estimation of $\mathbf{B}$ and $\mathbf{G}$ is done by minimizing a cost function $c(\mathbf{B},\mathbf{G})$, which is a weighted sum of three terms: a reconstruction error term $c_r(\mathbf{B},\mathbf{G})$, a temporal continuity term $c_t(\mathbf{G})$, and a sparseness term $c_s(\mathbf{G})$:

$$c(\mathbf{B},\mathbf{G}) = c_r(\mathbf{B},\mathbf{G}) + \alpha c_t(\mathbf{G}) + \beta c_s(\mathbf{G}), \quad (1)$$

where α and β are the weights for the temporal continuity term and sparseness term, respectively.

Figure 4 shows the result of applying the algorithm to a signal made up of two notes played by two different instruments (trumpet and oboe). Two components were calculated corresponding to the harmonic spectrum of the two notes played. Their temporal gain is showed in the two upper plots and they give information about when the notes are being played and its level over time.

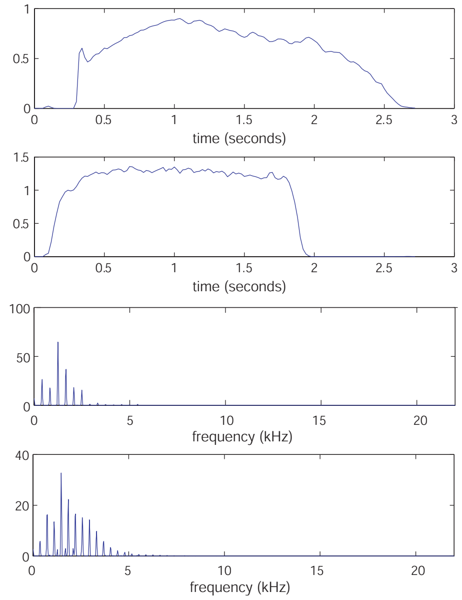


Figure 4 NMF components estimated from a mixture signal of two notes (trumpet and oboe). Gains are plotted on the top and basis functions at the bottom.

3.2. CASA speech segregation

The Monaural Speech Segregation Algorithm is a CASA algorithm. Specifically it is a system for voiced speech segregation. Pitch can be characterized by several perception mechanisms [19]. For resolved harmonics, the system generates segments based on temporal continuity and cross-channel correlation, and it groups them according to their periodicities. For unresolved harmonics, it generates segments based on common amplitude modulation (AM) in addition to temporal continuity and groups them accordingly. Pitch estimation of the target speech is an important step of the algorithm and determines the quality of the final segregated stream. Figure 5 shows the main stages of the algorithm.

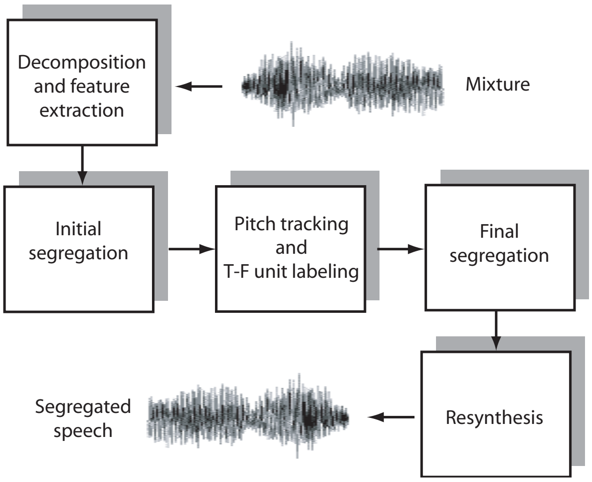

Decomposition and feature extraction
Mixture
Initial segregation
Pitch tracking and T-F unit labeling
Final segregation
Resynthesis
Segregated speech

Figure 5 Schematic diagram of the CASA multistage system for speech segregation.

3.3. Azimuth Discrimination and Resynthesis

The ADRess (Azimuth Discrimination and Resynthesis) exploits the use of the pan pot as a means to achieve image localizations within stereophonic recordings. In this kind of material, only an interaural intensity difference exists between left and right channels for a single source. It uses gain scaling and phase cancellation techniques to expose frequency dependent nulls across the azimuth domain, from which source separation and resynthesis is carried out. The right and left frequency-azimuth planes at a certain time-frame are given by:

$$Az_{R(k,t)} = \left| Lf(k) - g(i) \cdot Rf(k) \right|, \qquad (2)$$

$$Az_{L(k,t)} = \left| Rf(k) - g(i) \cdot Lf(k) \right|, \qquad (3)$$

where *Lf*(*k*) and *Rf*(*k*) are the magnitude spectrum of the left and right channels and *g*(i) is an azimuth gain factor. Figure 6 shows the azimuth domain for a stereo signal of a two partial mixture.

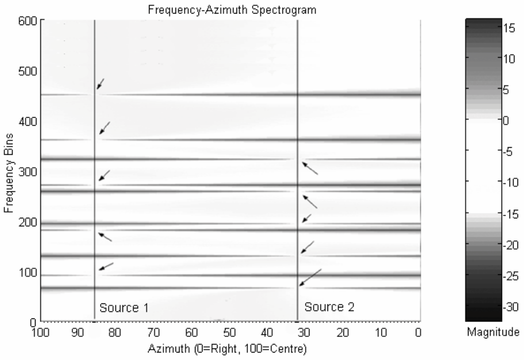


Figure 6 The Frequency-Azimuth spectrogram for a mixture of 2 synthetic sources each comprising of 5 non-overlapping partials. The arrows indicate frequency dependent nulls caused by phase cancellation. Extracted from [17].

4. SUBJECTIVE EVALUATION

4.1. Description of the experiments

As we can see, there are a lot of approaches to the problem of recovering sources from a mixture. They not only take profit of the characteristics of the signals involved, but in the characteristics of our auditory system or the signal acquisition set up. This wide range of algorithms brings to front another problem: how to evaluate the performance of a separation algorithm.

From an objective point of view, some criteria have been proposed in the literature, including Signal-to-distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifacts Ratio (SAR). Although they are related to the perceived audio quality in many cases, they do not model auditory phenomena of loudness weighting and spectral masking.

In [20] some guidelines for subjective evaluation of separation algorithms are given and an adaptation of the MUSHRA standard is proposed. In the WFS framework, subjective evaluation tests should take into account that spatial positioning of the sources is an additional parameter of interest. In order to evaluate the quality of resynthesized scenes with the separated sources, we have compared them with reference scenes created from originally separated signals, where spatial configuration of the sources is kept the same.

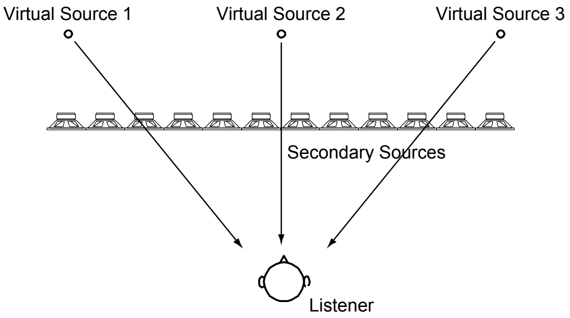


Figure 7 Scene configuration for subjective evaluation. Listener should perceive three different sources coming from left, center and right.

For this work, we have resynthesized several acoustic scenes by separation of monaural and stereo recordings. Three main case studies for different acoustic scenes have been proposed, in order to apply specific algorithms to specific mixtures. In the first two scenes, both monaural and stereo mixtures have been employed in the experiments. In the third one, only a monaural mixture has been tested.

- Scene 1 is a mixture made up of three ambient sounds: an ambulance (left), a car horn (right) and water dripping sound (front).

- Scene 2 is a music recording of a pop song made up of three sources: singing voice (front), piano (right) and drums (left).

- Scene 3 consists of the same ambulance and car horn of scene 2 (left) mixed up with male speech (right) ('It's time for some intervention here'). (monaural).

Subjective evaluation was carried out by means of listening tests. The test was performed employing a jury composed of 20 people. The subjects sit in front of the WFS array and the different scenes are presented to them successively.

First, an acoustic scene is presented using separated sources obtained from one of the implemented algorithms commented in section 3. Then, they are asked several questions related to the number of sources they can notice and also to the perception of their spatial location. After this, the scene is presented again with the original sources and they are asked the same questions. The last step of the test consists of evaluating the quality of the resynthesized scene by listening to it again once the original scene has been presented.

Specifically, three aspects are considered in relation to the test procedure previously commented:

1. Source identification: ability of identifying the number of sources present in the mixture.

2. Source localization: ability of identifying the direction of arrival of the sources.

3. Quality evaluation: subjective sound quality of the resynthesized scene in comparison to the reference one.

The allowed score for the third test is: excellent (5), good (3,75), fair (2,5), poor (1,25) and bad (0).

For the first and second tests, subjects' answers are given a score by comparing their answers in the scene composed using sources separated by means of the algorithm under test with their answers in the reference scene. The maximum score is always given when the answer in both cases (with separated sources and original sources) is the same. The final score for source identification and localization is the mean of the whole scores of the jury.

4.2. Results

Results of the tests are given in Table 1. It shows the score for the ADRess and NMF algorithms in case of ambient sound and music. Scene 3 was only processed with the CASA algorithm.

	Scene 1: Ambient sound			Scene 2: Music		
	Source identification	Source localization	Quality evaluation	Source identification	Source localization	Quality evaluation
ADRess	4.4	4.5	3.4	4.2	4.7	0.9
NMF	3.7	3.3	1.5	1.5	0.1	0.1

Table 1 Results for ADRess and NMF subjective evaluation in the WFS system

The NMF algorithm works worse than the ADRess algorithm both in the ambient scene and in the music scene. It must be taken into account that the ADRess algorithm takes profit of the stereo signal and the NMF algorithm works with a mono signal.

Moreover, it is interesting to appreciate the difference in subjective quality evaluated by the subjects between music and ambient sound. With the ADRess algorithm, subjective quality for Scene 1 (ambient sound) was 3.4, but 0.9 for Scene 2 (music). This difference shows how subjects tend to be more critical in their evaluation when music is being played, especially when singing voice is present. Source identification and localization is quite good for both algorithms in the ambient sound scene but poorer when the NMF algorithm is applied to music.

The CASA speech segregation algorithm was applied to Scene 3 in order to segregate speech from the other sounds in the mixture. Table 2 shows the scores obtained from the listening tests.

	Scene 3: Ambient + speech		
	Source identification	Source localization	Quality evaluation
CASA speech segregation	5	2.5	3.6

Table 2 Results for the CASA speech segregation algorithm

The monaural speech segregation algorithm was evaluated ‘good’ in terms of quality, but all of the subjects noticed that speech was not completely coming from a unique direction. This is because the algorithm was thought to segregate voiced speech (vowels), leaving as background fricative consonants and high-frequency components of speech. These background components disturbed the perception of speech making confusing its spatial location.

5. CONCLUSIONS

In this paper, one of the difficulties of the Wave-Field Synthesis systems to be widely deployed has been addressed. This difficulty resides in the fact that most of the commercial recorded material is in stereo format and there is no possibility to obtain the original multitrack recording.

The use of Sound Source Separation techniques to overcome this problem has been proposed in this paper, although existing algorithms are yet far from perfection resulting in audible artifacts.

First, a review of the different separation algorithms has been carried out. Next, some algorithms for sound source separation have been implemented and tested in order to resynthesize acoustic scenes in a WFS system. A subjective testing campaign involving a jury of 20 people has been carried out. The results show that the perceived subjective quality vary with the nature of the scene, being music more critical than ambient sound. The algorithms used in this work do not give high quality separated signals but masking effects in the WFS reproduction stage relax the quality needed in the separation if they are spatially mixed again.

Although the results are not definitive, they open a research line to work further.

6. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science under the project TEC2006-13883-C04-01 and FEDER funds.

7. REFERENCES

[1] A. J. Berkhout, “A Holographic Approach to Acoustic Control”, J. Audio Eng. Soc., 36, 977-995, 1988.

[2] A. J. Berkhout, D. de Vries, P Vogel, “Acoustic Control by Wave field Synthesis”, J. Acoust. Soc. Am., vol 93, pp. 2765-2778, 1993.

[3] C. Jutten , M. Babaie-Zadeh, “Source separation: principles, current advances and applications”, presented at the 2006 German-French Institute for Automation and Robotic Annual Meeting, IAR 2006, Nancy, France, November 2006..

[4] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation", IJIST (International Journal of Imaging Systems and Technology), 2005.

[5] K. Torkkola, "Blind separation for audio signals: are we there yet?", Proceedings of the Workshop on Independent Component Analysis and Blind Signal Separation, 1999.

[6] Li Y., Wang D.L, "Separation of singing voice from music accompaniment for monaural recordings", IEEE Transactions on Audio, Speech, and Language Processing*,* in press.

[7] E. Vincent, X. Rodet, A. Röbel, C. Févotte, É. Le Carpentier, R. Gribonval, L. Benaroya, and Fréderic Bimbot, "A tentative typology of audio source separation tasks", ICA 2003.

[8] C. K. Lee, D. G. Childers, "Cochannel speech separation", Journal of the Acoustical Society of America, 83(1), 1988.

[9] T. F. Quatieri, R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech", IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1), 1990.

[10] E. Vincen, X. Rodet, "Music transcription with ISA and HMM", in Proceedings of the 5th International Symposium on Independent Component Analysis and Blind Signal Separation, Granada, Spain, 2004.

[11] T. Virtanen, "Unsupervised Learning Methods for Source Separation", in Signal Processing Methods for Music Transcription, eds. Klapuri, A., Davy, M., Springer-Verlag, 2006.

[12] T. Virtanen, "Sound Source Separation in Monaural Music Signals", PhD. Thesis, presented at Tampere University of Technology, November 2006.

[13] T. Virtanen, "Accurate Sinusoidal Model Analysis and Parameter Reduction by Fusion of Components", presented at the 110th Audio Engineering Society Convention, Amsterdam, Netherlands 2001.

[14] S. Dubnov, "Extracting sound objects by independent subspace analysis", presented at the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland, June 2002.

[15] D.L. Wang, G. J. Brown. "Computational Auditory Scene Analysis: Principles, Algorithms, and Applications". IEEE Press/Wiley-Interscience, 2006.

[16] A. Bregman, "Auditory scene analysis". MIT Press, Cambridge, USA, 1990.

[17] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis". Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFTX 04), 2004.

[18] G. Hu, D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation". IEEE Transactions on Neural Networks, vol. 15, pp. 1135-1150, 2004.

[19] C. Michey, A. J. Oxenham, "Sequential F0 comparisons between resolved and unresolved harmonics: No evidence for translation noise between two pitch mechanisms", Journal of the Acoustical Society of America, Vol. 116, No. 5, pp. 3038–3050, November 2004.

[20] E. Vincent, M G. Jafari and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms". 2006 ICA Research Network Workshop, University of Liverpool, September 2006.

Comparación de técnicas de separación de voz en sistemas de videoconferencia avanzados

Máximo Cobos, José J. López, Andrés Cebrián, Alberto González

max_cob@iteam.upv.es, jjlopez@dcom.upv.es, anceblpe@teleco.upv.es, agonzal@dcom.upv.es

Instituto de Telecomunicaciones y Aplicaciones Multimedia, Universidad Politécnica de Valencia

Abstract — **Acoustic beamforming techniques use microphone arrays to achieve speaker enhancement and separation by means of spatial filtering . The use of such arrangement would provide many advantages in real communication systems and multimedia applications. These include advanced videoconference systems, speech recognition, speaker identification, source localization, hand-free communications or hearing aids. Array signal processing can enhance signals coming from a specific direction, whereas interfering signals coming from other directions are attenuated. This way, using spatial information given by a set of sensors, separating a signal from a mixture is possible. Unfortunately, the conventional theory is based on certain assumptions that are difficult to fulfil in practice: sensors must be accurately situated, sources must be in the far field and the signals generated by the sources must be of narrow band. Although many methods have been developed to address these problems, the use of several sensors can be also exploited by means of statistical techniques, specifically by Independent Component Analysis (ICA) algorithms. In this case, the assumptions are made in a statistical context: non-gaussian sources and statistical independence. Moreover, a suitable model for the mixtures must be taken into account. In this paper, voice signals picked up by a linear microphone array in a room are used to evaluate performances of both spatial and statistical techniques. Practical limitations of both methods are found out and compared for a real case using objective performance measures.**

I. Introducción

El sistema auditivo humano tiene la capacidad de identificar y separar señales en ambientes donde múltiples fuentes de sonido se mezclan con mayor o menor nivel. Este efecto se conoce en la literatura científica como "*cocktail party effect*", donde las señales de voz de diferentes individuos que hablan a la vez son extraídas automáticamente por el sistema auditivo humano. Esta capacidad humana todavía no ha sido alcanzada de forma tan perfecta por ningún sistema automático, pero existen bastantes progresos en el campo y sus aplicaciones en el campo de las telecomunicaciones son múltiples: mejora de los sistemas de reconocimiento de voz, sistemas de sonido 3D, videoconferencia de alto realismo, audífonos, etc.

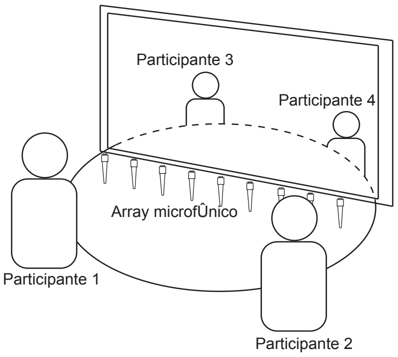


Fig. 1. Sistema de videoconferencia avanzado. Los participantes son detectados y realzados entre ambos extremos de la conexión.

En el escenario de aplicación de la Figura 1, se muestra un sistema de videoconferencia avanzado, donde existen dos grupos de oyentes separados por lo que se conoce como ventana visual y acústica. En este caso se desea conseguir un sistema que mediante un conjunto de micrófonos [1] sea capaz de extraer y realzar la voz de cada uno de los oyentes individualmente. Dentro de este contexto de separación de fuentes, dos de las técnicas más interesantes para ser empleadas son: la conformación de haz mediante arrays de micrófonos (*beamforming*) y el *análisis de componentes independientes* (ICA).

Las técnicas de beamforming consiguen la separación mediante el principio de filtrado espacial. El filtrado espacial tiene como objetivo realzar las señales que provienen de una dirección específica mediante la configuración apropiada de un array de micrófonos, al mismo tiempo que atenúa las señales interferentes provenientes de otras direcciones del espacio [2].

La otra técnica propuesta, que ha despertado mucho interés en los últimos años, es el análisis de componentes independientes. Los algoritmos de ICA están basados puramente en la estadística de las señales a separar, lo que ha permitido encontrar aplicación en numerosas áreas, como la biomedicina o las telecomunicaciones. El modelo de señal más utilizado es el de mezcla instantánea. Sin embargo, en el caso de señales de voz recogidas dentro de una sala el modelo de mezcla convolutiva es más realista, aunque como veremos en el punto III, esto también complica bastante los algoritmos a utilizar así como la implementación de las soluciones.

En este artículo nos centraremos en evaluar la calidad de la separación de voz de dos personas hablando simultáneamente utilizando ambas técnicas: un array de micrófonos lineal y un algoritmo de ICA para mezclas convolutivas. El análisis se realizará mediante parámetros objetivos y se discutirá la posible implementación en sistemas de videoconferencia reales.

II. SEPARACIÓN POR FILTRADO ESPACIAL O BEAMFORMING

Un array es una agrupación de elementos (emisores o receptores), dispuestos según una determinada configuración geométrica, con objeto de mejorar un sistema de comunicaciones en comparación al que se obtendría con un sistema de un solo elemento (emisor o receptor). En general, cuanto mayor tamaño (apertura) tenga el elemento receptor con relación a la longitud de onda λ, mayor será su directividad y por tanto su capacidad de rechazo de señales no deseadas. La directividad depende de dos características: del tamaño del array y de la frecuencia de trabajo. Cuanto mayor es la frecuencia, menor es λ, por lo que el array es mayor en términos de longitud de onda y es por tanto más directivo.

El beamformer más simple es el llamado de "retardo y suma", el cual suma las señales de cada micrófono en fase para la dirección de interés, cancelando las señales de otras direcciones. Los llamados beamformers adaptativos intentan cancelar fuentes no deseadas adaptando los pesos utilizados con el tiempo o bien mediante un proceso de entrenamiento. En general, un beamformer adaptativo formado por *L* micrófonos es capaz de eliminar únicamente *L-1* fuentes de ruido diferentes.

Las características propias de la señal de voz y el camino de transmisión acústica hacen que los resultados y conceptos habitualmente manejados para señales radioeléctricas no sean de aplicación directa en el campo del sonido. En concreto:

1) La señal de voz tiene un ancho de banda relativo muy amplio (5 octavas en voz, 10 en audio en general). Un ancho de banda relativo amplio hace que la directividad del array varíe mucho en su margen de frecuencias.
2) La longitud de onda de las frecuencias más bajas de la voz son del orden de varios metros, por lo que se requerirían arrays enormes para conseguir una directividad aceptable en estas frecuencias.
3) En un escenario típico de captación de voz existen muchas reflexiones en la sala que introducen ecos y reverberación.
4) La distancia fuente/array puede variar mucho en relación al tamaño del mismo y situarse en campo cercano.

La tarea del beamforming consiste básicamente en filtrar la señal obtenida por cada micrófono por un filtro FIR, para posteriormente sumar cada canal y producir una única salida y(t).

$$y(t) = \sum_{i=1}^{I} w_i\,(t) * x_i(t), \tag{1}$$

donde *I* es el número de micrófonos, $x_i(t)$ es la señal eléctrica captada por cada micrófono y w_i(t) es la respuesta impulsiva del filtro utilizado en el canal *i*. La convolución con cada filtro puede verse en el dominio transformado como una multiplicación por un coeficiente para cada frecuencia, calculando la FFT de la señal captada por cada micrófono.

Los pesos óptimos que maximizan la potencia de una fuente situada en (r_0, θ_0, φ_0) vienen dados por:

$$\underline{w}(r_0, \theta_0, \varphi_0) = \frac{\underline{a}(r_0, \theta_0, \varphi_0)}{\underline{a}^H(r_0, \theta_0, \varphi_0)\underline{a}(r_0, \theta_0, \varphi_0)}, \tag{2}$$

donde $a(r_0, \theta_0, \varphi_0)$ representa el *steering vector* (o vector de direcciones de llegada). Este vector incluye principalmente el efecto del camino acústico entre la fuente y el array, que se traduce en atenuación y cambio de fase o retardo para cada uno de los micrófonos. Para el caso de un array lineal uniforme (ULA) orientado en el eje *z* y con su centro en el origen, los pesos que maximizan la potencia de señal con dirección de llegada θ_0 son:

$$W_i(\omega, \theta_0) = \frac{1}{I} e^{\,j\frac{\omega}{c} z_i \cos\theta_0}, \tag{3}$$

siendo z_i la posición del micrófono i-ésimo y *c* la velocidad del sonido (aproximadamente 340 m/s). La dependencia con la frecuencia *ω* de los pesos en cada micrófono se soluciona trabajando con la FFT de las señales de entrada. De esta forma, el problema de filtrado en cada canal se reduce a la multiplicación por un coeficiente distinto para cada raya espectral de la FFT de la señal recogida por cada micrófono.

Un array se puede considerar como un muestreo espacial de las ondas que llegan a él. De forma análoga al muestreo temporal, el aliasing también puede aparecer en este caso en forma de lóbulos no deseados de la misma amplitud que el lóbulo principal en direcciones no deseadas en el diagrama de directividad. Para el caso peor (*endfire*), la separación entre micrófonos no debe ser superior a la mitad de la longitud de onda (correspondiente a la frecuencia máxima). Otro problema es la dependencia con la frecuencia de la directividad, siendo ésta peor para baja frecuencia. La solución a este problema no es siempre viable en la práctica, pues para conseguir una directividad aceptable en baja frecuencia se requieren arrays enormes.

III. Análisis de Componentes Independientes para mezclas convolutivas

La separación de fuentes mediante métodos de ICA, consiste en estimar un conjunto de señales fuente $s_i(n)$ usando la información de las señales de mezcla u observaciones $x_j(n)$ en cada uno de los canales, de la forma:

$$x_j(n) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \quad (j=1,...,M), \tag{4}$$

donde s_i es la señal de la fuente i, x_j es la señal recibida por el micrófono j, y h_{ji} es la respuesta impulsiva de P coeficientes del canal entre la fuente i y el micrófono j. Este es el modelo que más se ajusta a una situación real, y se conoce como mezcla convolutiva. En el caso de mezclas instantáneas, el modelo es más sencillo, pues h_{ji} son escalares y modelan únicamente una diferencia de ganancia entre cada fuente y cada sensor. A pesar de la calidad de separación obtenida por las técnicas de ICA en el caso de mezclas instantáneas (válidas en muchas aplicaciones), conseguir resultados similares en mezclas convolutivas es en la actualidad un problema por resolver. La gran mayoría de algoritmos de ICA que trabajan sobre este modelo lo hacen en el dominio de la frecuencia y es un problema en el que se sigue trabajando intensivamente en la actualidad.

Para recuperar las señales, se deben estimar los filtros $w_{ij}(k)$ de Q coeficientes, de forma que se consiga una estimación de cada fuente $y_j(n)$ de la forma:

$$y_i(n) = \sum_{j=1}^{M}\sum_{q=1}^{Q} w_{ij}(p)x_j(n-q+1) \quad (i=1,...,N) \tag{5}$$

La teoría de ICA se basa precisamente en estimar estos filtros de forma que las señales recuperadas sean mutuamente independientes. En este artículo consideraremos el caso de dos micrófonos y dos fuentes ($N=M=2$). En el caso de mezclas convolutivas la estimación de estos filtros presenta cierta complejidad. Trasladar el problema al dominio frecuencial convierte la mezcla convolutiva en problema de mezcla instantánea. Utilizando una representación STFT de T puntos:

$$\mathbf{X}(\omega,m) = \mathbf{H}(\omega)\mathbf{S}(\omega,m) \tag{6}$$

donde ω es un punto de frecuencia de la transformada y m es el número de ventana temporal de la STFT. El vector de fuentes es $\mathbf{S}(\omega,m) = [\mathrm{S}_1(\omega,m), \mathrm{S}_2(\omega,m)]^{\mathrm{T}}$ y $\mathbf{X}(\omega,m) = [\mathrm{X}_1(\omega,m), \mathrm{X}_2(\omega,m)]^{\mathrm{T}}$ es el vector de observaciones. $\mathbf{H}(\omega)$ es una matriz de mezcla 2×2 invertible. El proceso de separación se puede escribir como:

$$\mathbf{Y}(\omega,m) = \mathbf{W}(\omega)\mathbf{X}(\omega,m) \tag{7}$$

donde $\mathbf{Y}(\omega,m) = [\mathrm{Y}_1(\omega,m), \mathrm{Y}_2(\omega,m)]^{\mathrm{T}}$ es el vector de fuentes estimadas y $\mathbf{W}(\omega)$ es una matriz de separación 2×2 para el punto de frecuencia ω. La teoría de ICA se basa en encontrar la matriz $\mathbf{W}(\omega)$ que hace a $\mathrm{Y}_1(\omega,m)$ e $\mathrm{Y}_2(\omega,m)$ independientes.

Aplicar este modelo en frecuencia introduce un problema importante: la ordenación de filas en la matriz $\mathbf{W}(\omega)$ que da lugar a la extracción de componentes independientes en la frecuencia ω es arbitraria. Como consecuencia, las distintas fuentes se recuperan con un orden diferente de frecuencias. Además de este problema, la ambigüedad de escala (también trivial en el dominio del tiempo) supone versiones filtradas de las componentes encontradas.

IV. Experimentos y Resultados

Para la evaluación de las dos técnicas comentadas en los anteriores apartados, se ha propuesto utilizar la configuración experimental mostrada en la Figura 2. El array lineal utilizado para la captación de voz consta de 8 micrófonos separados entre si 5 cm y situado a una distancia de 1.3 m del suelo. Dos altavoces con posiciones angulares 0º y 90º reproducen simultáneamente señales de voz diferentes con la misma potencia. El experimento se ha realizado en una sala acondicionada acústicamente ($T_{60} \approx 0.2$ s) para minimizar la influencia de los ecos en el experimento.

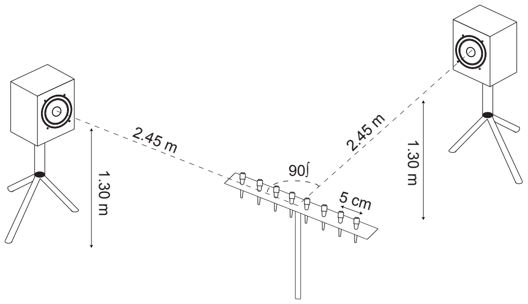
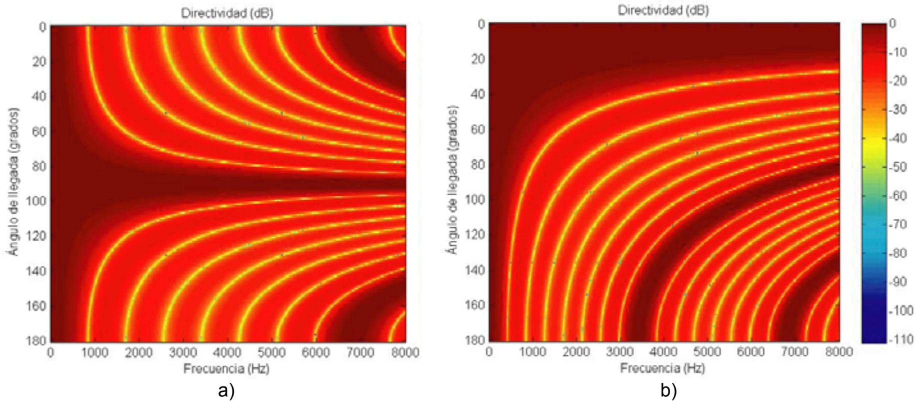


Fig. 2. Montaje experimental para la captación de señales de voz con un array de 8 micrófonos.

Las señales reproducidas por los altavoces son una voz masculina (0º) y otra femenina (90º). Los pesos utilizados para el procesado de las señales captadas por cada uno de los micrófonos se calcularon siguiendo la ecuación (2). En la Figura 3 se muestran los diagramas de directividad resultantes de los pesos utilizados para la separación de cada una de las fuentes. En la Figura 3.a se puede observar cómo para todas las frecuencias la dirección de máxima ganancia es 90º y en la Figura 3.b la dirección de máxima ganancia siempre es 0º. Desafortunadamente, los diagramas de directividad no son constantes con la frecuencia. De hecho, a mayor frecuencia, el array es más grande en términos de λ, por lo que es más directivo. En baja frecuencia, sucede lo contrario y el diagrama se vuelve prácticamente omnidireccional. Como consecuencia de ello, el array no conseguirá discriminar bien las señales no deseadas en baja frecuencia. Esta característica junto con la de aliasing espacial (aparecen direcciones con igual ganancia a partir de cierta frecuencia), suponen las mayores limitaciones de la técnica utilizada.

Para la separación de las dos señales de voz mediante ICA, se ha utilizado el algoritmo descrito en [3] utilizando únicamente la señal de los dos micrófonos centrales del array. Este algoritmo utiliza una técnica que pretende encontrar aquellas fuentes que sean lo menor gaussianas posible mediante la maximización de la negentropía. El problema de permutabilidad de puntos de frecuencia en las componentes independientes se soluciona en este algoritmo reordenando según mínima distancia entre puntos consecutivos de frecuencia para la matriz $\mathbf{W}(\omega)$. El algoritmo también compensa el problema de escalado y agiliza la convergencia aprovechando la información espacial de las fuentes. Todas las señales se grabaron con una frecuencia de muestreo de 44 kHz y se remuestrearon a 16 kHz para la comparación.



a) b)

Fig. 3. Diagramas de directividad para la separación de fuentes situadas en a) 90º y b) 0º.

La evaluación objetiva de la calidad obtenida con la separación está bien discutida en [4]. Se proponen cuatro parámetros o medidas basados en la definición habitual de la SNR con algunas modificaciones. Estas cuatro medidas son el SDR (*Signal to Distortion Ratio*), el SIR (*Signal to Interference Ratio*) y el SAR (*Signal to Artifacts Ratio*). Todas ellas han sido medidas siguiendo [5]. Se puede observar cómo de forma general, los resultados son mejores utilizando ICA que beamforming, aunque el tiempo de computación es también mucho mayor. La medida que tiene más en cuenta el realzado de un participante sobre otro es el SIR, que para la técnica ICA es más elevado que el SDR y el SAR, sobretodo para la voz masculina. El resto de medidas también son en general mejores para la voz masculina que para la femenina.

Respecto al coste computacional, el tiempo necesario para la separación por beamforming para una trama de 3 s es de aproximadamente 0.2 s, siendo mucho menor que el tiempo consumido por el algoritmo ICA, que es de 36.5 s. Ambas pruebas se hicieron en un ordenador Intel® Core™ 2 1.86 Ghz.

TABLA I
EVALUACIÓN DE AMBAS TÉCNICAS

	SDR (dB)		SIR (dB)		SAR (dB)	
	Masculina	Femenina	Masculina	Femenina	Masculina	Femenina
Beamforming	4.5	-0.8	6.9	0.2	8.9	8.7
ICA	3.9	1.6	22.9	11.2	4.0	2.4

VII. Conclusión

En este artículo se han evaluado de forma objetiva los resultados obtenidos en la separación de dos voces mediante beamforming y análisis de componentes independientes para mezclas convolutivas. Ambas técnicas han sido introducidas brevemente y aplicadas a un caso real utilizando un array lineal de 8 micrófonos. Podemos concluir de los resultados obtenidos, que la separación de las señales de voz ha sido mejor con el algoritmo de ICA utilizado que con el beamforming tradicional, sin embargo el alto coste computacional del algoritmo de ICA puede hacer inviable su aplicación en sistemas en tiempo real.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Programa Nacional de I+D+I del Ministerio de Educación y Ciencia con el proyecto TEC2006-13883-C04-01 y fondos FEDER.

REFERENCIAS

[1] F. Khalil, J. P. Jullien, A. Gilloire, "Microphone Array for Sound Pickup in Teleconference Systems," *JAES* Volume 42 Number 9 pp. 691-700; September 1994.

[2] T. L. Tung, K. Yao, D. Chen, R. E. Hudson, and C. W. Reed, "Source Localization and Spatial Filtering Using Wideband Music and Maximum Power Beamforming for Multimedia Applications," *in Proc. IEEE SiPS, Oct.* 1999, pp. 625-634.

[3] R. Prasad, H. Saruwatari, A. Lee, and K. Shikano, "A Fixed-Point ICA Algorithm for Convoluted Speech Signal Separation," *in ICA2003 4th Int. Symp.*, pp. 579-584, Nara, Japan, April 2003.

[4] E. Vincent, R. Gribonval, C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. on Speech and Audio Procc*, vol. 14, nº4, pp. 1462-1469, 2006

[5] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide," *IRISA Technical Report* 1706., Rennes, France, April 2005. http://www.irisa.fr/metiss/bss_eval/.

Stereo Audio Source Separation based on Time-Frequency Masking and Multilevel Thresholding

Máximo Cobos *, José J. López [1]

Audio and Communications Group (GTAC), Insitute of Telecommunications and Multimedia Applications (iTEAM), Technical University of Valencia, Spain

Abstract

Source separation and up-mixing in real commercial music recordings is a challenging problem. In the last years, some algorithms have provided interesting results, but the problem remains unsolved. In this paper we describe a method for automatic separation of the sources present in a two channel mixture based on the panning coefficients used in the stereo mixdown. The sources are separated by estimating time frequency masks using the multilevel extension of the Otsu's thresholding algorithm used in image segmentation. A refinement step is also carried out for extraction and reassignment of inter-source residuals. Examples of application and performance evaluation are also discussed.

Key words: Sound source separation, stereo music mixtures, multilevel thresholding
PACS: 43.60.+d, 42.72.+q

1. Introduction

In the last years, source separation and up-mixing techniques applied to music recordings have received increased attention [1][2][3]. Generally, when trying to extract different sources from a mixture, many assumptions have to be made, not just about the sources but also about the mixing process as well. In the source separation framework, the number of mixture channels, the number of sources, and the nature of the mixing process are key issues to be taken into account [4]. Independent component analysis and beamforming techniques use several sensors to extract different sources from a set of mixtures channels. The problem becomes more difficult when the number of sources increase or the mixture is convolutive [5]. Single sensor source separation methods for music mixtures are also a common topic in recent publications [6][7][8].

When dealing with stereo commercial music recordings, only the information in the left and right channels can be exploited and the mixture is generally underdetermined, which means that there are more sources than mixture channels. The type of mixing process roughly categorizes stereo recordings into studio recordings and live recordings. Studio recordings are made by instantaneous mixing of recorded mono or stereo tracks which usually correspond to different sources. These tracks are mixed using amplitude panning to create the stereophonic effect. Mono or stereo reverberation can be added artificially in the mix.

In this paper we present a method for automatic separation of sources in stereo instantaneous mixtures. As other methods used in stereo source separation, the framework used in this paper is

* Corresponding author.
Email addresses: macoser1@iteam.upv.es (Máximo Cobos), jjlopez@dcom.upv.es (José J. López).
[1] Address: Audio and Communications Signal Processing Group (GTAC), iTEAM, Camino de Vera S/N, Building 8G, Valencia, Spain. (+34 616 25 13 95)

also based on the analysis of the interaural intensity difference (IID) existent between the two observation channels in the STFT domain [9][10]. A basic assumption made by these algorithms is that in the time-frequency transform domain, signal components corresponding to different sources do not overlap significantly [11]. This is often called the W-disjoint orthogonality assumption. Whereas some separation methods need specific information about the panning configuration [2] or human attendance [3] for completing the separation process, the method described in this work performs an automatic estimation of the optimum time-frequency masks for different sources. A $\log^{-1}$ weighted histogram and the multilevel extension of the Otsu's thresholding algorithm [12] are used for this purpose.

2. Stereo Mixing Model

Studio recordings can be modelled as a sum of J amplitude panned sources $s_j(t)$, $j = 1 \ldots J$ convolved with reverberation impulse responses $r_i(t)$ for each channel $i = 1, 2$ [2]. The stereo mixture channels can be written as

$$x_i(t) = \left[\sum_{j=1}^{J} a_{ij} s_j(t) \right] * r_i(t) \quad i = 1, 2 \tag{1}$$

where a_{ij} are the amplitude panning coefficients used in the stereo mixdown. Assuming a short reverberation impulse response in each channel, the mixture becomes instantaneous:

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t) \quad i = 1, 2. \tag{2}$$

There are many ways to set the panning coefficients in the analog mixers and Digital Audio Workstations. Most of them use the sinusoidal energy-preserving panning law [3], based in the pan knob $\phi \in [0, 1]$

$$a_{1j} = \cos\left(\frac{\phi\pi}{2}\right) \tag{3}$$

$$a_{2j} = \sin\left(\frac{\phi\pi}{2}\right) \tag{4}$$

$$a_{1j}^2 + a_{2j}^2 = 1. \tag{5}$$

To formalize, we denote the STFT's of the channel signals $x_i(t)$ as $X_i(k, m)$, where k is the frequency index and m is the time index. Given the linearity of the STFT, we can write the STFT of each channel as

$$X_i(k, m) = \sum_{j=1}^{J} a_{ij} S_j(k, m) = \sum_{j=1}^{J} S_{ij}(k, m) \tag{6}$$

where $S_{ij}(k, m) = a_{ij} S_j(k, m)$ is the image of source j in channel i in the STFT transform domain.

The approach taken in this paper works independently for sources panned to different azimuth sectors. A source s_j is said to be panned to the left if $a_{1j} > a_{2j}$. If $a_{1j} < a_{2j}$ the source is said to be panned to the right. If $a_{1j} = a_{2j}$ we say that the source is panned to the center. The mixing model can be also written as

$$x_i(t) = \sum_{p=1}^{J_1} a_{ip} s_p(t) + \sum_{q=1}^{J_2} a_{iq} s_q(t) + \sum_{c=1}^{J_c} a_{ic} s_c(t) \tag{7}$$

where J_1 is the number of sources panned to the left, J_2 the number of sources panned to the right and J_c the number of sources panned to the center.

3. Pan Map in Quasi W-Disjoint Orthogonal Sources

Speech mixtures have shown to be well approximated by the W-disjoint orthogonality assumption [10]. However, in practice, when music mixtures are considered, this assumption is not as well anymore. The separation method described in this paper is based on the W-disjoint orthogonality assumption, and therefore there will be always an error when estimating the sources from their mixing coefficients. In this section, the pan map of a stereo mixture is introduced and the deviation error from the W-disjoint orthogonality case is studied.

3.1. *W-Disjoint Orthogonal Sources*

If the mixing coefficients are time invariant, the amplitude ratio between the left and right channels for a single source remains constant:

$$\frac{s_{1j}(t)}{s_{2j}(t)} = \frac{a_{1j}}{a_{2j}} \tag{8}$$

The sources are said to be W-disjoint orthogonal if they do not overlap in the STFT transform domain. This can be expressed mathematically [4] as

$$S_i)(k, m) S_j(k, m) = 0 \quad \forall i \neq j, \forall k, m \tag{9}$$

Thus, only an active source will be present in each time-frequency point, and the ratio between the magnitude of the STFT of the mixture channels, $\rho(k,m) = |X_1(k,m)|/|X_2(k,m)|$, will correspond to the ratio between the mixing coefficients of the active source, given by

$$\rho(k,m) = \rho_W(k,m) = \frac{|S_{1a}(k,m)|}{|S_{2a}(k,m)|} = \frac{a_{1a}}{a_{2a}} \tag{10}$$

where the subindex W refers to the W-disjoint orthogonality assumption and a is the index of the active source in the time-frequency point (k,m).

We define the pan map as the logarithm of $\rho(k,m)$ and it represents the log-mixing ratio of each time-frequency point in the STFT transform domain:

$$P(k,m) = 20\log\left(\rho(k,m)\right) \tag{11}$$

3.2. *Pan Map Deviation*

The mixing ratio would uniquely identify the time-frequency components of the sources in the stereo mix only when they are all panned to different locations and do not overlap significantly in the transform domain, as discussed in [2]. In practice, the sources present in the audio signal (and especially in music recordings) are overlapped in time and frequency. This means that there will be a set C of interfering sources which have energy in a shared time-frequency point with a main source of interest s_j:

$$\rho(k,m) = \frac{|S_{1j}(k,m)| + \sum\limits_{i\in C}|S_{1i}(k,m)|}{|S_{2j}(k,m)| + \sum\limits_{i\in C}|S_{2i}(k,m)|} \tag{12}$$

The estimated mixing ratio for the source of interest will correspond to the mixing ratio under the W-disjoint orthogonality assumption plus a deviation produced by the interfering sources:

$$\begin{aligned}\rho(k,m) &= \frac{|S_{1j}(k,m)|}{|S_{2j}(k,m)|} \\ &\quad + \frac{\sum\limits_{i\in C}|S_{1i}(k,m)| - \frac{|S_{1j}(k,m)|}{|S_{2j}(k,m)|}\sum\limits_{i\in C}|S_{2i}(k,m)|}{|S_{2j}(k,m)| + \sum\limits_{i\in C}|S_{2i}(k,m)|} \\ &= \rho_W(k,m) + \Delta \end{aligned} \tag{13}$$

Taking the logarithm of $\rho(k,m)$, we can write the pan map as

$$\begin{aligned} P(k,m) &= 20\log\left(\rho(k,m)\right) = 20\log\left(\rho_W(k,m)\right) \\ &\quad + 20\log\left(1 + 10^{(\log\Delta - \log(\rho_W(k,m)))}\right). \end{aligned} \tag{14}$$

As we can see in the above equation, the pan map $P(k,m)$ can be decomposed into two terms, one corresponding to the pan map under the W-disjoint orthogonality assumption and another one corresponding to a deviation produced by interfering sources (i.e. non W-disjoint orthogonal sources). Note that $P(k,m)$ will be positive for that sources panned to the left channel that have a small Δ. Sources panned to the right will correspond to negative points in $P(k,m)$.

Figure 1 shows the deviation error for three pan positions of the source of interest.. The deviation is represented as a function of the Signal-to-Interference Ratio (SIR) and the pan position of the interfering sources. For a fixed SIR, the maximum error is debt to the source panned to the most distant location. In general, as the interfering energy increases, the deviation error is increased too. The smallest error corresponds to sources panned to the center whereas sources panned to azimuth edges have a severe deviation error.

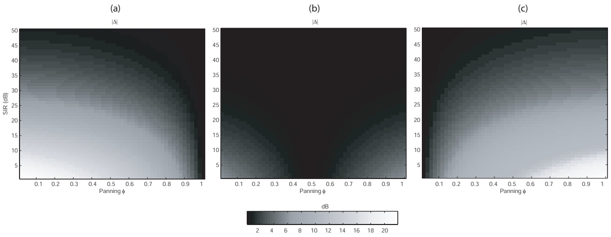


Fig. 1. Mixing ratio deviation error. (a) Source of interest panned to the top left. (b) Source of interest panned to the center. (c) Source of interest panned to the right.

4. Separation Framework

Most of the stereo separation methods consist in estimating the coefficients used in the mixing process in order to make a clustering of time frequency points that have a similar mixing ratio [9][2][3][10]. The general approach is to define a two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures. Peaks corresponding to the relative attenuation and delay mixing parameters of each source are observed and time-frequency masks are formed for each peak, allowing the separation. This approach has shown to provide good results when dealing with two channel speech mixtures, but insufficient when music recordings are considered. When multiple instruments and singing voice are present, the overlap is much more significant and the mixing ratio varies so much that no clear peaks can be observed in the histogram.

In this paper we describe a method based on the analysis of the pan map and time-frequency masking. The developed method begins with a split of the pan map into two different azimuth panned regions which will be similarly processed. We propose to use a perceptual weighted histogram made up with time-frequency points only in the medium frequency range, where the sources are supposed to concentrate their energy. Then, the Fast Multilevel Otsu Algorithm [13] is used for searching the optimum thresholds that maximize the between-class variance of the mixing ratio values. These thresholds will define a set of binary masks that will be later reassigned to different sources. A refinement step is also carried out for removing inter-source residuals.

4.1. *Pan Map Splitting*

The first step for the separation of the sources consists in splitting the pan map $P(k,m)$ into two parts corresponding to sources panned to the left and sources panned to the right. This is made by creating two binary masks, one for positive values of the pan map and another one for the negative values

$$U^{(1)}(k,m) = \begin{cases} 1 & if \quad P(k,m) \geq 0 \\ 0 & if \quad P(k,m) < 0 \end{cases} \tag{15}$$

$$U^{(2)}(k,m) = \begin{cases} 1 & if \quad P(k,m) \leq 0 \\ 0 & if \quad P(k,m) > 0 \end{cases} \tag{16}$$

Multiplying the pan map by these masks, we split $P(k,m)$ into two parts

$$\begin{aligned} P(k,m) &= \sum_{i=1}^{2} P(k,m)U^{(i)}(k,m) \\ &= \sum_{i=1}^{2} P^{(i)}(k,m) \end{aligned} \tag{17}$$

Figure 2 shows the mixture spectrograms $X_i(k,m)$, where four sources (piano, guitar, drums and singing voice) are overlapped in the time-frequency transform domain. The original stereo is a instantaneous mixture of these sources sampled at 44.1 kHz. The STFT was carried out using a Hanning window of length 180 ms with 75% overlap. The pan map and the two binary masks obtained for these music mixtures are represented in Figure 3.

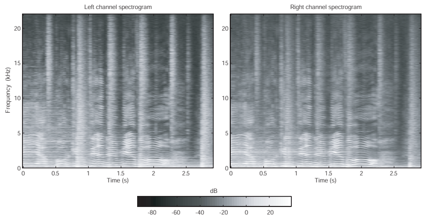


Fig. 2. Left channel and right channel amplitude spectrograms.

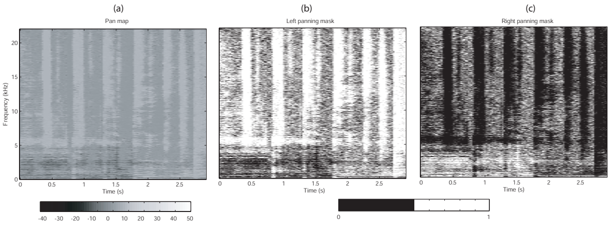


Fig. 3. (a) Pan map obtained from the spectrograms showed in Figure 2. (b) Binary mask for sources panned to the left. (c) Binary mask for sources panned to the right.

4.2. *Histogram Formation*

The second step consists in estimating the mixing ratios of the sources panned to the left and those panned to the right separately by analyzing the absolute value of the previously calculated pan maps $|P^{(i)}(k,m)|$.

First, $|P^{(1)}(k,m)|$ and $|P^{(2)}(k,m)|$ are normalized, obtaining $Pn^{(1)}(k,m)$ and $Pn^{(2)}(k,m)$. Then, two histograms of L uniform containers in the range $[0,1]$ are formed for $Pn^{(i)}(k,m)$, just taking into account only points in a medium frequency range $[k_{min}, k_{max}]$. The center of each container is given by

$$z_n = \frac{1}{2L}(2n+1) \quad n = 0 \ldots L-1 \tag{18}$$

We take k_{max} the index of the closest frequency to 4 kHz and k_{min} the index of the closest frequency to 100 Hz. This histogram is calculated as a $\log^{-1}$ weighted sum of the number of points that lie in each of the L containers previously defined. This procedure gives a greater value to points in the lower part of the frequency range of interest:

$$H(n) = \sum_{i \in n} g(k_i), \tag{19}$$

where $g(k_i)$ is the weighting factor for a point (k_i, m_i) with value $Pn^{(1)}$ or $Pn^{(2)}$ (depending on the channel considered) in the value range defined by container n. This is a first approximation to perceptual weighting and can be calculated as

$$g(k) = \frac{\log(100)}{\log(100 + k - k_{min})}. \tag{20}$$

Figure 4 shows the histograms obtained for $Pn^i(k, m)$ from the pan map shown in Figure 3. In the original mixdown, guitar was panned to the right, drums and piano were panned to the left, and singing voice was positioned at the centre of the azimuth plane. If the sources were perfectly W-disjoint orthogonal, clear peaks corresponding to each source should be easily identified in the histogram containers corresponding to the different mixing ratios of the present sources. In [2] a range of values where the mixing ratios for each source may vary are selected using a Gaussian window. However, the central point of the window must be specified for carrying out the separation. In [3] a human-assisted criterion is used. We propose to use a multilevel thresholding algorithm for selecting the range of values in the histograms corresponding to each source. This way, the sources are extracted automatically by maximizing their inter-class variance defined by Otsu.

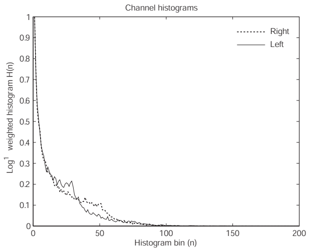


Fig. 4. Weighted histogram for left and right channels.

4.3. *Multilevel Thresholding*

Thresholding is an important technique for image segmentation which is used for identification and extraction of targets from its background on the basis of the distribution of pixel intensities in image objects. In our separation framework, image segmentation and source extraction from a mixture pan map can be observed from the same point of view. In the separation context, we try to find the thresholds that maximize the inter-class variance of the distribution of mixing ratios over the time-frequency transform domain.

We briefly describe the Otsu's algorithm [12]:

The probability of the $\log^{-1}$ weighted mixing ratio in the middle of container n of the histogram is given by:

$$p_n = \frac{H(n)}{N} \tag{21}$$

where $N = \sum_{n=1}^{L} H(n)$.

In the case of bi-level thresholding, the time-frequency points are divided into two classes, c_1 with mixing ratios in the range given by the histogram bins $n \in [1, \ldots, t]$ and c_2 with values within the bins $n \in [t+1, \ldots, L]$. Then, the probability distributions for the two classes are:

$$c_1 : p_1/\omega_1(t), \ldots, p_t/\omega_1(t) \tag{22}$$

$$c_2 : p_{t+1}/\omega_2(t), p_{t+1}/\omega_2(t), \ldots, p_L/\omega_2(t) \tag{23}$$

where $\omega_1(t) = \sum_{n=1}^{t} p_n$ and $\omega_2(t) = \sum_{n=t+1}^{L} p_n$.

The means for classes c_1 and c_2 are

$$\mu_1 = \sum_{n=1}^{t} n \frac{p_n}{\omega_1(t)} \tag{24}$$

$$\mu_2 = \sum_{n=t+1}^{L} n \frac{p_n}{\omega_2(t)} \tag{25}$$

Let μ_T be the mean mixing ratio for the whole image. Then:

$$\omega_1 \mu_1 + \omega_2 \mu_2 = \mu_T \tag{26}$$

$$\omega_1 + \omega_2 = 1 \tag{27}$$

Otsu defined the between-class variance as:

$$\sigma_B^2 = \omega_1 (\mu_1 - \mu_T)^2 + \omega_2 (\mu_2 - \mu_T)^2 \tag{28}$$

For bi-level thresholding, Otsu verified that the optimal threshold t^* is chosen so that the between-class variance σ_B^2 is maximized, that is:

$$t^* = \text{Arg max} \left\{ \sigma_B^2(t) \right\} \quad 1 \leq t \leq L \tag{29}$$

The previous formula can be easily extended to multilevel thresholding. Assuming that there are $M - 1$ thresholds, $\{t_1, t_2, \ldots, t_{M-1}\}$, which divide the original pan map into M classes: c_1 for $[1, \ldots, t_1]$, c_2 for $[t_1 + 1, \ldots, t_2]$, ..., c_i for $[t_{i-1} + 1, \ldots, t_i]$ and c_M for $[t_{M+1} + 1, \ldots, L]$, the optimal thresholds

$t_1^*, t_2^*, \ldots, t_{M-1}^*$ are chosen by maximizing σ_B^2 as follows:

$$\{t_1^*, t_2^*, \ldots, t_{M-1}^*\} = \text{Arg max}\left\{\sigma_B^2(t_1, t_2, \ldots, t_{M-1})\right\} \tag{30}$$

where $1 \leq t \leq L$ and

$$\sigma_B^2 = \sum_{k=1}^{M} \omega_k(\mu_k - \mu_T)^2, \tag{31}$$

with

$$\omega_k = \sum_{n \in c_k} p_n \tag{32}$$

$$\mu_k = \sum_{n \in c_k} n \frac{p_n}{\omega(k)} \tag{33}$$

The ω_k in Eq. 32 is regarded as the zeroth-order cumulative moment of the kth class c_k, and the numerator in Eq. 33 is regarded as the first-order cumulative moment of the kth class c_k, that is

$$\mu(k) = \sum_{n \in c_k} n p_n. \tag{34}$$

Regardless of the number of classes being considered during the thresholding process, the sum of the cumulative probability functions of M classes equals one, and the mean of the mixing ratios considered is equal to the sum of the means of M classes weighted by their cumulative probabilities. The between-class variance in Eq. 31 can thus be rewritten as

$$\sigma_B^2 = \sum_{k=1}^{M} \omega_k \mu_k^2 - \mu_T^2. \tag{35}$$

Because the second term in Eq. 35 is independent of the choice of the thresholds, the optimal thresholds can be chosen by maximizing $\sigma_B'^2$, which is defined as the summation term on the right-hand side of Eq. 35:

$$\sigma_B'^2 = \sum_{k=1}^{M} \omega_k \mu_k^2 \tag{36}$$

A faster algorithm can be achieved by recursive calculation of Eq. 36. [13]. Let us define the look-up tables for the $u - v$ interval:

$$P(u,v) = \sum_{n=u}^{v} p_n \tag{37}$$

$$S(u,v) = \sum_{n=u}^{v} n p_n \tag{38}$$

For index $u = 1$, equations 37 and 38 can be rewritten as

$$P(1, v+1) = P(1,v) + p_{v+1} \tag{39}$$

and $P(1,0) = 0$

$$S(1, v+1) = S(1,v) + (v+1)p_{v+1} \tag{40}$$

and $S(1,0) = 0$.

From equations 39 and 40, it follows that

$$P(u,v) = P(1,v) + P(1,u-1) \tag{41}$$

and

$$S(u,v) = S(1,v) + S(1,u-1) \tag{42}$$

Now, the modified between-class variance $\sigma_B'^2$ can be rewritten as

$$\sigma_B^2 = G(1,t_1) + G(t_1+1,t_2) + \\ \ldots + G(t_{M-1}+1, L), \tag{43}$$

where the modified between-class variance of class c_i is defined as

$$G(t_{i-1}+1, t_i) = \frac{S(t_{i-1}+1, t_i)^2}{P(t_{i-1}+1, t_i)}. \tag{44}$$

The search range for the maximal $\sigma_B'^2$ is $1 \leq t_1 \leq L - M + 1, t_1 + 1 \leq t_2 \leq L - M + 1, \ldots, t_{M-1} + 1 \leq t_{M-1} \leq L - 1$.

The final thresholding values will be those in the middle of containers $n = t_i^*$:

$$Th_i = z_n|_{n=t_i^*} \tag{45}$$

4.4. *Binary Masking*

Once the optimum thresholds have been calculated for $Pn^{(i)}$, we are able to define the binary masks corresponding to each class.

Let's call $Th_i^{(1)}$ and $Th_i^{(2)}$ the optimum thresholds for sources panned to the left and right channels, respectively. Figure 5 shows the optimum thresholds found for a case where three different classes are considered in each channel histogram. Note that we can search for an arbitrary number of classes in each channel, even if the number of sources panned to that channel is lower. When this happens, a refinement step for clustering several classes to a same source must be carried out. This will be further discussed in the next step.

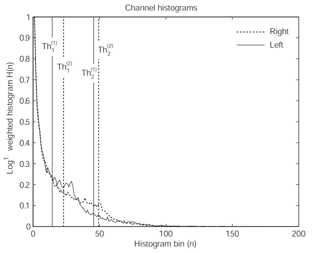


Fig. 5. Optimum thresholds.

The binary masks for sources panned to the left, are given by

$$U_i^{(1)} = \begin{cases} U^{(1)} & \text{if} \quad Th_{i-1}^{(1)} < Pn^{(1)} \leq Th_i^{(1)} \\ 0 & \text{elsewhere} \end{cases} \tag{46}$$

with $i = 1 \dots M_1$, being M_1 the number of classes to be estimated in the left channel histogram, $Th_0^{(1)} = 0$ and $Th_{M_1}^{(1)} = 1$.

Similarly, for the right channel:

$$U_i^{(2)} = \begin{cases} U^{(2)} & \text{if} \quad Th_{i-1}^{(2)} < Pn^{(2)} \leq Th_i^{(2)} \\ \quad 0 & \text{elsewhere} \end{cases} \tag{47}$$

with $i = 1 \dots M_2$, being M_2 the number of classes to be estimated in the right channel histogram, $Th_0^{(2)} = 0$ and $Th_{M_2}^{(2)} = 1$.

Figure 6a and 6b shows the masks formed by applying the obtained thresholds to $Pn^{(1)}$ and $Pn^{(2)}$, respectively.

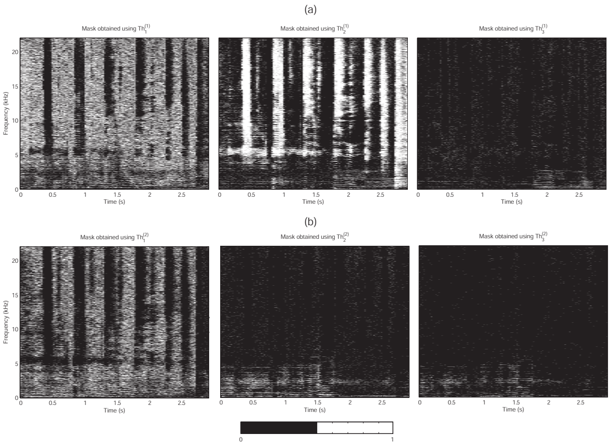


Fig. 6. Primary binary masks. (a) Masks obtained by applying the thresholds in the first histogram to the left binary mask. (b) Masks obtained by applying the thresholds in the second histogram to the right binary mask.

4.5. *Class Reassignment*

As already stated, there's no restriction in the multilevel thresholding process for defining the number of classes M_i in $Pn^{(i)}$. This means that, if the number of sources panned to the left (or right) is lower than the number of classes defined in the thresholding step ($M_i > J_i$), then more than one mask may correspond to the same source. Independently of the number of classes considered, when a source is panned to the center, there will be always two masks corresponding to that source: $U_1^{(1)}$ and $U_1^{(2)}$. In fact, two masks corresponding to a same source are always azimuth adjacent, which simplifies the reassignment step.

First, the masks are azimuth ordered to form a single set of masks $\mathbf{B} = \{B_i, B_2, \dots, B_{M_1+M_2}\} = \left\{U_{M_1}^{(1)}, \dots, U_1^{(1)}, U_1^{(2)}, \dots, U_{M_2}^{(2)}\right\}$.

Although many ways for comparing binary images can be used, we propose a simple way for doing it just by taking a $N \times M$ grid for each mask B_i and computing the number of non-zero points m_n in each cell. This way, a vector for each mask $\mathbf{m}_i = [m_1\ m_2 \dots m_{N\times M}]^T$ is formed. Then, we calculate the mean distance between all the adjacent vectors $\mathbf{m}_i$ and $\mathbf{m}_{i+1}$:

$$d_{i,i+1} = \frac{1}{N \times M} \sum_{n=1}^{N\times M} |\mathbf{m}_i(n) - \mathbf{m}_{i+1}(n)| \tag{48}$$

for $i = 1, \dots, M_1 + M_2 - 1$.

If $d_{i,i+1}$ is a local or absolute minimum of the whole distances sequence, then their corresponding masks are added: $B'_i = B_i \cup B_{i+1}$. After this reassignment step, a set of $J' \leq M_1+M_2$ different masks are available for retrieving the original sources. In Figure 7, the reassigned masks ordered in azimuth (from left to right) are shown.

4.6. *Source image retrieval*

We can estimate the source images in each channel just applying the calculated masks to the STFT of each channel, conserving the phase information of the mixture:

$$\hat{S}_{ij}(k,m) = |X_i(k,m)| B'_j \, e^{j\angle X_i} \tag{49}$$

for $j = 1 \dots J'$ and $i = 1, 2$.

The estimated sources in time domain will be

$$\hat{s}_j = \text{STFT}^{-1}\left\{S_{1j} + S_{2j}\right\} \tag{50}$$

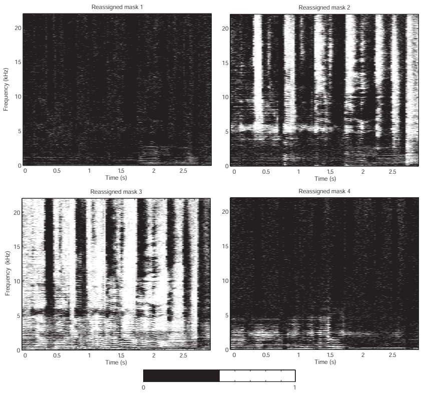


Fig. 7. Masks obtained after the reassignment step.

for $j = 1 \dots J'$.

4.7. *Refinement Step*

A refinement step can be carried out for reassigning inter-source residuals in the separated sources by applying the described method recursively. For a separated source $\hat{S}_{ij}(k,m)$, we can calculate its normalized pan map as

$$Pn_j^{(1)} = Pn^{(1)} B'_j \quad \text{or} \quad Pn_j^{(2)} = Pn^{(2)} B'_j \qquad (51)$$

depending on if the source is panned to the left or to the right. A histogram for Pn_j is carried out as explained in Subsection 4.2. Next, a bilevel thresholding ($M = 2$) is applied to the pan map, segregating the initial mask B'_j into two masks, as in Subsections 4.3 and 4.4. At this time, we may have two masks, one of them corresponding to the primary source in $\hat{S}_{ij}(k,m)$, and another one corresponding to a residual of one of the separated sources adjacent to the one one considered: $\hat{S}_{i(j-1)}(k,m)$ or $\hat{S}_{i(j+1)}(k,m)$.

The reassigning procedure described in Subsection 4.5 can be used for adding this residual to the more similar adjacent mask. Figure 8a shows the final masks obtained after the refinement step. For comparing purposes, the optimum masks that can be achievable under the W-disjoint orthogonality assumption are shown in Figure 8b. These masks have been obtained with a priori knowledge of the original sources. The optimum mask for a given source is formed by creating an all-zero mask and setting to one only those time-frequency points where the energy of the source is greater than the energy of the other sources.

After this reassigning step, the final estimated sources are recovered by applying this masks to the STFT of the mixture channels and calculating the inverse STFT of the result. We show the obtained waveforms in Figure 9a, and the original waveforms of the sources in Figure 9b. The similarity between the separated sources and the original is obvious.

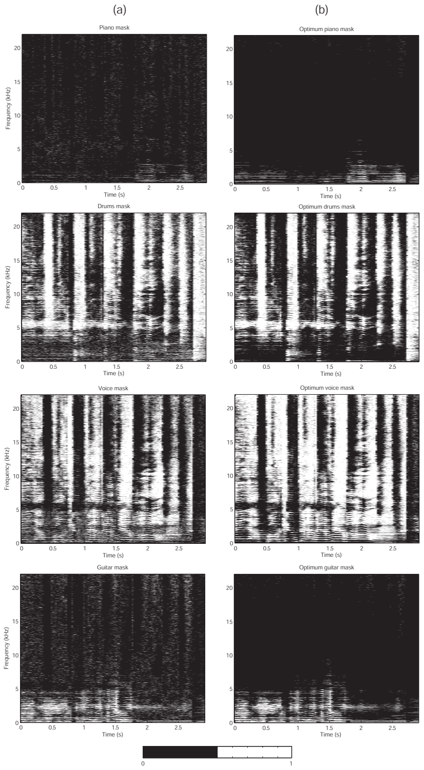


Fig. 8. (a) Final masks. (b) Optimum masks assuming W-disjoint orthogonality obtained from the original sources.

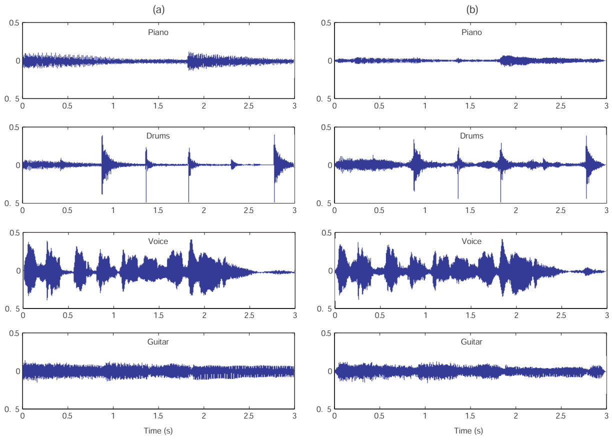


Fig. 9. Waveform results. (a) Separated sources. (b) Original sources.

5. Practical Considerations

Although the above separation framework can be extended to an arbitrary number of sources, a practical limit is always present when applying the described processing to a stereo music mixture. This is again a consequence of the mixing process and the W-disjoint orthogonality assumption, which is far from being true for audio sources (for speech mixtures is a more realistic assumption). Note that if two different sources are mixed with the same pan, then they will be extracted as a unique source, as they both will have the same mixing ratio. Moreover, if many sources are present although panned to different azimuth positions, as each time-frequency point is assigned to a different source, the recovered sources will be plenty of artifacts due to the non linear filtering process. This makes not very useful to search for more than three classes in each normalized pan map ($M_1 = M_2 = 3$).

6. Performance Evaluation

Some performance measures in source separation processing have already been described in the literature. In [14] the objective performance evaluation of sound source separation algorithms is well discussed. In some applications it may be relevant to allow more or less distortions, not necessarily related to the theoretical indeterminacies of the problem. The evaluation procedure developed in [14] takes into account the application for which a given separation algorithm is oriented. For example, in musical applications, it may be important to recover the sources up to a simple gain since arbitrary filtering modifies the timbre of musical instruments. The assumptions made for applying the performance criteria are:

– the true source signals are known,

– a family of allowed distortions is chosen

The computation of the criteria involves two successive steps. First, $\hat{s}_j$ is decomposed as

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif}, \qquad (52)$$

where $s_{target} = f(s_j)$ is a version of s_j modified by an allowed distortion $f \in \mathcal{F}$, and where e_{interf}, e_{noise} and e_{artif} are respectively the interferences, noise and artifacts error terms. From this decomposition, some numerical performance criteria are defined. The Source to Distortion Ratio

$$\mathrm{SDR} := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}, \qquad (53)$$

Table 1
Objective measures with time invariant gain distortion allowed

	SDR		SIR		SAR	
$\hat{s}_j$	ADRess	MLTS	ADress	MLTS	ADRess	MLTS
Piano	-0.2	-2.7	28.4	15.7	-0.2	-2.5
Drums	5.1	3.8	19.6	11.2	5.3	5.0
Voice	-2.3	10.4	13.1	20.6	-2.0	10.9
Guitar	0.1	4.2	13.0	16.9	0.6	4.5

the Source to Interferences Ratio

$$\mathrm{SIR} := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2}, \qquad (54)$$

the Sources to Noise Ratio

$$\mathrm{SNR} := 10 \log_{10} \frac{\|s_{target} + e_{interf}\|^2}{\|e_{noise}\|^2}, \qquad (55)$$

and the Sources to Artifacts Ratio

$$\mathrm{SAR} := 10 \log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2}, \qquad (56)$$

These measures can be interesting for comparing several algorithms. Given a family of allowed distortions, the SIR and SAR are valid as performance measures regarding two separate goals: rejection of the interferences and absence of forbidden distortions or “burbling” artifacts. The SNR is valid as a measure of rejection of the sensor noise. The SDR can be seen as a global performance measure.

In this paper, we have compared the ADRess algorithm for music separation [1] with the described multilevel thresholding separation method (MLTS), using the same window length and overlap values (180 ms and 0.75%). For that purpose, we have used the MATLAB toolbox *BSS_EVAL* [15], distributed online under the GNU Public License. Table 1 shows the SDR, SIR and SAR for the estimated sources using both algorithms when only a time invariant gain distortion is allowed. Table 2 shows the evaluated performance when a time invariant filtering is considered (128 taps allowed in the distortion filter). SNR was not considered because no noise was assumed. As we can see in the tables, similar results are obtained allowing both distortions.

The results show how the source panned to the center (voice) is the one extracted with the higher SIR and SAR, as it was expected from the study of the deviation error in Section 3. This source and the source panned to the right (guitar) are better extracted using the MLTS method than the ADRess

Table 2
Objective measures with time invariant filtering distortion allowed

	SDR		SIR		SAR	
$\hat{s}_j$	ADRess	MLTS	ADress	MLTS	ADRess	MLTS
Piano	-0.2	-2.1	19.8	10.9	0.3	-1.5
Drums	5.4	3.9	18.5	10.3	5.7	5.4
Voice	-1.9	10.5	11.1	19.4	-1.4	11.2
Guitar	0.4	4.8	11.9	16.9	1.0	5.2

algorithm, which presents periodic gain artifacts. These artifacts are debt to the fact that no cancellations are found in time frames where the source has little energy, producing a noise gate effect. In the MLTS method, the residuals in the extracted sources make the listening more comfortable and this would probably affect positively in subjective evaluation tests. The evaluated tracks can be listened to at http://personales.upv.es/macoser1.

7. Summary

In this paper we presented a new method for stereo audio source separation based in the multilevel version of the Otsu's thresholding algorithm used in image segmentation. The thresholds are calculated so that the interclass variance of a map formed by different mixing ratio values is maximized. A perceptual weighting approach is also carried out in the histogram formation step before searching for the optimal thresholds. A set of binary masks are obtained for each source after a reassignment and refinement step.

As many other separation algorithms, the source signals are assumed to be W-disjoint orthogonal. Although music signals don't satisfy this assumption at all, this can be enough for many applications in music information retrieval, remixing purposes or sound scene resynthesis.

In the present work, an example has been carried out through the different separation steps. The example considered was a fragment of a pop song where three instruments and singing voice were mixed instantaneously. The results were compared with the ADRess algorithm using the *BSS_EVAL* toolbox for objective evaluation of blind source separation algorithms. Sources nearly panned to the center showed very good values of SIR and SAR. Further work need to be done for improving the quality of separated tracks by combining other separation methods or making use of the phase information, which has not been considered in this paper.

8. Acknowledgements

This work was supported by the Ministry of Education and Science under the project TEC2006-13883-C04-01 and FEDER funds.

References

[1] D. Barry, B. Lawlor, E. Coyle, Sound source separation: Azimuth discrimination and resynthesis, in: 7th Conference on Digital Audio Effects (DAFTX 04), 2004.

[2] C. Avendano, Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2003.

[3] M. Vinyes, J. Bonada, A. Loscos, Demixing commercial music productions via human-assisted time-frequency masking, in: Audio Engineering Society 120th Convention, Paris, France, 2006.

[4] E. Vincent, M. Jafari, S. Abdallah, M. Plumbey, M. Davies, Blind audio source separation, C4dm-tr-05-01, Queen Mary, University of London (November 2005).

[5] R. Prasad, H. Saruwatari, A. Lee, K. Shikano, A fixed-point ica algorithm for convoluted speech signal separation, in: ICA2003 4th International Symposium, 2003, pp. 579–584.

[6] G. Siamantas, M. Every, E. Szymanski, Separating sources from single-channel musical material: a review and future directions, in: Digital Music Research Network Conference 2006, London, England, 2006.

[7] V. T., Sound source separation in monaural music signals, Ph.D. thesis, Tampere University of Technology (November 2006).

[8] Y. Li, D. Wang, Separation of singing voice from music accompaniment for monaural recordings, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 1475–1487.

[9] A. Jourjine, S. Richard, O. Yilmaz, Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00, Vol. 5, Turkey, 2000, pp. 2985–2988.

[10] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Transactions on Signal Processing.

[11] P. O'Grady, B. Pearlmutter, S. Rickard, Survey of sparse and non-sparse methods in source separation, International Journal of Imaging Systems and Technology (IJIST).

[12] N. Otsu, A threshold selection method from gray-level histogram, IEEE Transactions on System Man Cybernetics SMC-9 (1) (1979) 62–66.

[13] P. Liao, T. Chen, P. Chung, A fast algorithm for multilevel thresholding, Journal of Information Science and Engineering 17 (2001) 713–717.

[14] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Transactions on Speech and Audio Processing 14 (4) (2006) 1462–1469.

[15] C. Févotte, R. Gribonval, E. Vincent, Bss_eval toolbox user guide, Tech. Rep. 1706, IRISA, Rennes, France (April 2006).