# PERFORMANCE ANALYSIS OF A SOURCE SEPARATION ALGORITHM

*Matthias Baeck and Udo Zölzer*

Department of Signal Processing and Communications
University of the German Federal Armed Forces Hamburg
Udo.Zoelzer@unibw-hamburg.de

## ABSTRACT

Source separation is an attractive preprocessing step for applying digital audio effects to a single source inside a signal mix. We present a performance analysis of a source separation algorithm based on time-frequency processing and its application to digital audio effects. The performance analysis gives insight to the main analysis parameters for the detection of the number of source signals inside the signal mix. We also analyze the main design parameters for the demixing operation which extracts a single source out of the signal mix.

## 1. INTRODUCTION

The ability of the human auditory system to extract a signal out of a mix of multiple signals is termed as the cocktail party effect. Source separation is a signal processing technique to build the so-called cocktail-party processor [1]. Based on the two microphone arrangement with two sound sources shown in Fig. 1 the signal at microphone 1 is $x_1(t) = a_{11}s_1(t-d_{11}) + a_{21}s_2(t-d_{21})$ and and the signal at microphone 2 is $x_2(t) = a_{12}s_1(t-d_{12}) + a_{22}s_2(t-d_{22})$. A perfect source separation algorithm should deliver an estimation of the number of sources in the signal mix, the angle of arrival and the distance to each source signal. For the general case [4, 5] the discrete-time microphone signals $x_1(n)$ and $x_1(n)$ can be written as

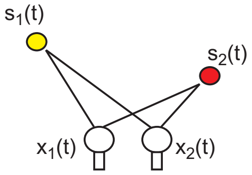

Figure 1: *Source separation of two signals arriving at two microphones.*

$$x_1(n) = \sum_{j=1}^{N} s_j(n) \tag{1}$$

$$x_2(n) = \sum_{j=1}^{N} a_j s_j(n-d_j), \tag{2}$$

where $s_j(n) = 1, \ldots, N$ are the $N$ signal sources, $d_j$ is the delay difference of signal source *j* between the two microphones, and $a_j$ is the relative amplitude of signal source *j* at microphone 2 referring to microphone 1.

## 2. SOURCE SEPARATION

A variety of algorithms for source separation have been reported in the literature. We will concentrate on combined time and frequency domain based approaches reported in [1, 2, 3, 4, 5]. The stereo signal is transformed into the time-frequency domain by a phase vocoder implementation discussed in [6]. Figure 2 shows the processing stages for a phase vocoder based source separation algorithm. The short-time Fourier transform of both microphone signals leads to

$$X_1(k,l) = \sum_{n=0}^{L-1} x_1(n) \cdot w(n-l) W_L^{kn} \tag{3}$$

$$X_2(k,l) = \sum_{n=0}^{L-1} x_2 \cdot w(n-l) W_L^{kn} \tag{4}$$

$$\text{with} \qquad W_L = e^{-j\frac{2\pi}{L}}$$

where $L$ is the size of the FFT (Fast Fourier Transform), $k$ is the frequency index and $l$ is the time index. The hop size for the short-time FFT is $L/4$. Based on the $X_1(k,l)$ and $X_2(k,l)$ we can compute the relative amplitude $a(k,l)$ and the delay difference $d(k,l)$. We will briefly explain the algorithm suggested in [4, 5]. Due to the discrete Fourier transforms (DFT) given by

$$\begin{aligned} \text{DFT}\{\sum_{j=1}^{N} s_j(n)\} &= \sum_{j=1}^{N} S_j(k), \\ \text{DFT}\{a \cdot s(n)\} &= a \cdot S(k), \qquad \text{and} \\ \text{DFT}\{s(n-d_j)_L\} &= S(k) W_L^{kd_j}, \end{aligned}$$

we can write (1) and (2) in the frequency domain as

$$X_1(k,l) = \sum_{j=1}^{N} S_j(k,l) \tag{5}$$

$$X_2(k,l) = \sum_{j=1}^{N} a_j S_j(k,l) W_L^{kd_j}, \tag{6}$$

or in matrix notation

$$\left[\begin{array}{c} X_1(k,l) \\ X_2(k,l) \end{array}\right] = \left[\begin{array}{ccc} 1 & \ldots & 1 \\ a_1 W_L^{kd_1} & \ldots & a_N W_L^{kd_N} \end{array}\right] \left[\begin{array}{c} S_1(k,l) \\ \vdots \\ S_N(k,l) \end{array}\right]. \tag{7}$$

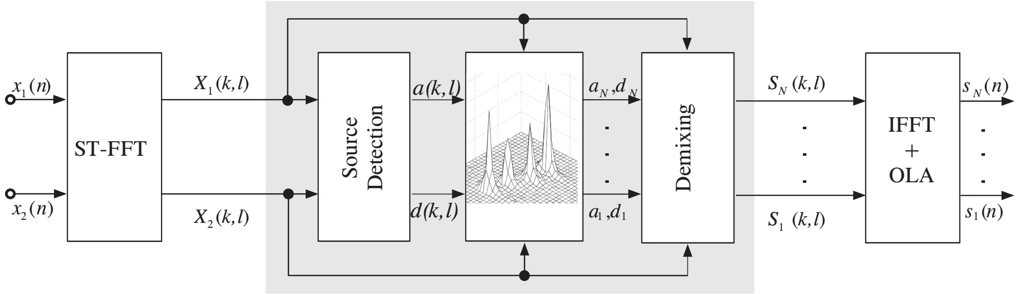

Figure 2: *Phase vocoder implementation of a source separation algorithm: estimation of signals inside a stereo signal and extraction of signals from the stereo signal.*

Source separation is possible, if the source signals are orthogonal in the time-frequency representation. The W-disjoint orthogonality defined in [4, 5] is given by

$$S_j(k,l) \cdot S_i(k,l) = 0 \qquad j,i = 1,\ldots,N \qquad i \neq j. \tag{8}$$

Every point $(k,l)$ on the time-frequency grid (spectrogram) corresponds to only one source. This means that in (7) only one term is valid. Accordingly we can write (7) as

$$\begin{bmatrix} X_1(k,l) \\ X_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j W_L^{kd_j} \end{bmatrix} S_j(k,l), \tag{9}$$

where index *j* represents one single source signal, which is active at point $(k,l)$ on the time-frequency grid.

Based on Eq. (9) every point $(k,l)$ on the time-frequency grid belongs to one pair $(a_j, d_j)$, which corresponds to the source signal $j$. By a simple division

$$\frac{X_2(k,l)}{X_1(k,l)} = a_j W_L^{kd_j} = a_j e^{-j\frac{2\pi k d_j}{L}} \tag{10}$$

we can derive the parameters $a_j$ und $d_j$ as follows:

$$\Rightarrow a_j = \left|\frac{X_2(k,l)}{X_1(k,l)}\right| \tag{11}$$

$$\begin{aligned} -\frac{2\pi k d_j}{L} &= \angle\frac{X_2(k,l)}{X_1(k,l)} \\ \Rightarrow d_j &= -\frac{L}{2\pi k}\angle\frac{X_2(k,l\tau_0)}{X_1(k,l)} \\ &= \frac{L}{2\pi k}\angle\frac{X_1(k,l)}{X_2(k,l)}. \end{aligned} \tag{12}$$

Computation of (11) and (12) for all $(k,l)$ leads to a time-frequency grid for the pairs

$$(a(k,l), d(k,l)) = \left(\left|\frac{X_2(k,l)}{X_1(k,l)}\right|, \frac{L}{2\pi k}\angle\frac{X_1(k,l)}{X_2(k,l)}\right). \tag{13}$$

With these two parameters of the time-frequency grid at frequency point *k* and time position *l* we can calculate a two-dimensional histogram $h(a,d)$, as described in [4, 5]. Dominant source signals inside a signal mix appear as peaks in the two-dimensional histogram $h(a,d)$. The corresponding parameters $a_j$ and $d_j$ for one detected signal are used for the extraction of this signal from the signal mix with the help of the time-frequency representation and a subsequent IFFT and overlap-add procedure (see Fig. 2).

## 3. PERFORMANCE ANALYSIS

The main applications of source separation for digital audio effects may be found in the areas of post-processing separated sources out of two-channel mixes. The source separation algorithm is tested with the following signals

- stereo signals based on two-microphone signals,
- artificial stereo signals based on amplitude and delay panning,
- and stereo signals based on amplitude panning.

### 3.1. Estimation of Source Signals

#### *3.1.1. Two-microphone signals*

Stereo signals based on two microphone recordings fit the assumptions for the source signals given in [4, 5]. The histogram should show the real parameters $a(k,l)$ and $d(k,l)$ which are needed for the demixing algorithm. Our analysis based on a number of stereo recordings shows that the sources mostly can be found as separated peaks in the histogram $h(a,d)$. Figure 3 shows an example for the histogram of two sources recorded by two microphones. One axis shows the $\log a$ values and the other axis represents the delay difference $d$ in samples. The recordings were made in an almost anechoic environment. The FFT length is $L = 2048$, the hop size is $L/4 = 512$ and grid size of the histogram $h(a,d)$ is 30 × 30. Figure 4 shows the amplitude/delay histogram for a stereo drum recording, where three peaks are clearly detectable. The highest peak belongs to the snare drum which is positioned in the left stereo image. The bass drum appears as two smaller peaks. Experiments with a stereo microphone arrangement show that the number of speakers is limited to three to four simultaneously speaking persons in close distance to the microphones.

#### *3.1.2. Artificial stereo signals*

Artificial stereo mixtures based on amplitude and delay panning are the most natural reproductions of real stereo recordings by two microphones. Figure 5 shows a histogram for a stereo signal containing two source signals which are positioned in the stereo mix by amplitude and delay panning. The peaks in the histogram are clearly detectable and much less spread around the real parameters $a$ and $d$, because there is obviously no room effect in the stereo mix in contrast to real recordings. Unfortunately, mixing consoles for multi-channel applications only offer amplitude panning for a

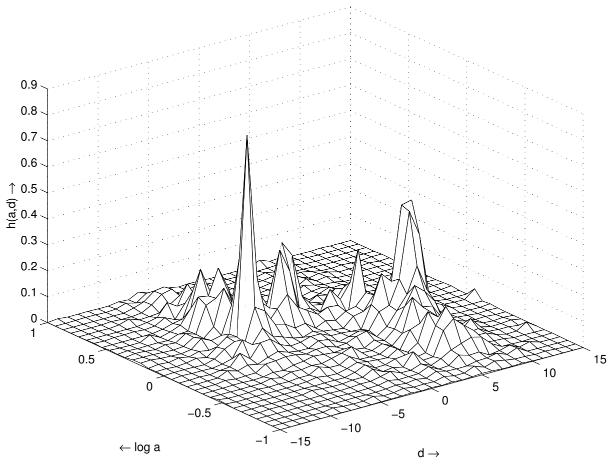

Figure 3: *Amplitude/delay histogram for two sources recorded by two microphones in an almost anechoic environment. The microphone distance is 50 cm and both speakers are are 50 cm in front of each microphone.*

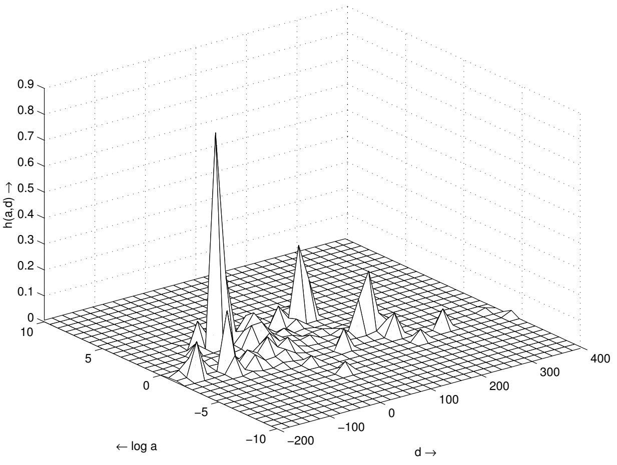

Figure 4: *Amplitude/delay histogram for stereo drum recording.*

single source signal. If a combination of amplitude and delay panning would be offered, the resulting stereo mix could be used to separate and extract the included source signals.

### 3.1.3. Stereo signals based on amplitude panning

Most artificial stereo signals are mixed by amplitude panning of several source signals. This results in a histogram showing peaks only on the amplitude-axis, i.e. with zero delay. The estimation of the parameters $a(k,l)$ and $d(k,l)$ via the histogram only works for a few sources which are positioned separately in the room. For an orchestra with many instruments close together, it is very difficult to detect separated peaks in the histogram.

### 3.1.4. Problems with delay estimation

Figure 6 shows a histogram for two sources with constant amplitude ratio but with the delay spread for one source signal. The condition for a correct determination of the delay difference is given by $|\omega d| < \pi$. This implies that the delay difference calculation is only limited to low frequencies. So the peak for one source in the histogram is spread along the delay difference. This effect limits the accuracy for the detection of the correct parameters to demix the source signals.

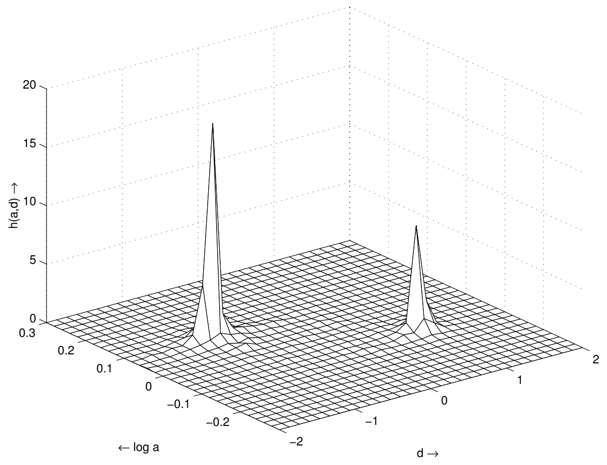

Figure 5: *Amplitude/delay histogram of artificial stereo mix with amplitude and delay panning.*

### 3.1.5. Problems due to peak spreading

Due to [5] the source signals are approximately orthogonal in the time-frequency domain. Our analysis shows that this approximation results in a peak spreading in the histogram. An example also depicted in Fig. 6 shows that the second source signal is spread around the correct parameters.

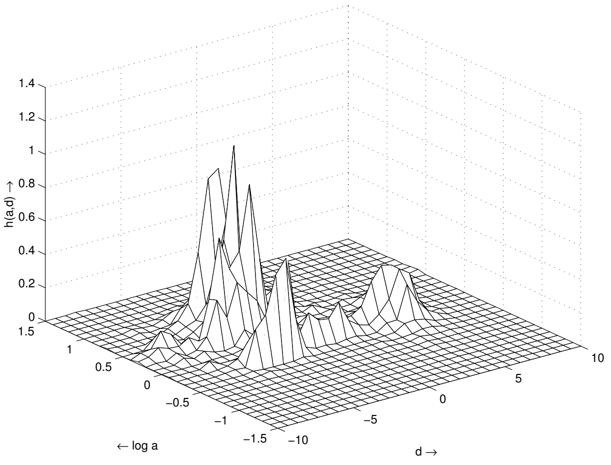

Figure 6: *Amplitude/delay histogram with delay spreading of one source signal and peak spreading of the other source.*

### 3.2. Demixing of Source Signals

Once we have found the corresponding parameters for one detected signal, we can extract one single source signal by the algorithm given in [4, 5]. Due to peak spreading it is necessary to use not only the parameters $a(k,l)$ and $d(k,l)$ where the highest peak is detected, but also the surrounding region of the peak in the histogram. Varying results of the demixing operation make a fine tuning of the region around the peak very important. Figure 7 shows the left and right signals of a stereo drum signal and depicts the spectrogram of the right drum signal. The amplitude/delay histogram is shown in Fig. 4. Separation in the time domain is of course easier to establish due to the time domain orthogonality of the involved drum signals. Figure 8 shows the extracted and normalized snare drum signal and its corresponding spectrogram. The result shows the effectiveness of the separation and demixing operation performed by the phase vocoder implementation of the source separation algorithm illustrated in Fig. 2.

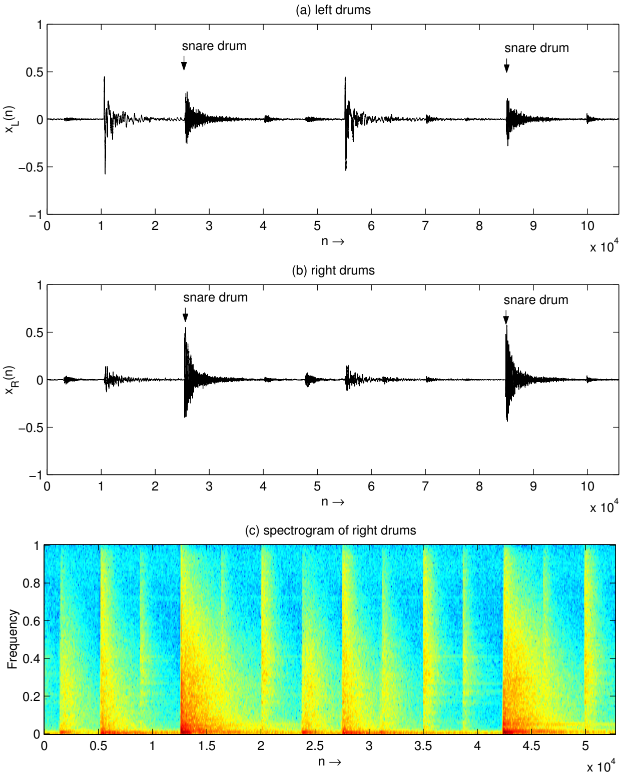

Figure 7: *Stereo drum signal and spectrogram of right signal.*

Stereo mixes, where the sources are very wide spread, are very difficult for source separation. This is the case when echo effects lead to virtual sources or the peak spreading is a result of further source signals with similar amplitude and delay parameters. The orthogonality in time and frequency is no longer valid in such situations. In such cases it is necessary to include a very wide region around the highest peak, which results in undesirable sources audible in the extracted signal. On the other hand, if sources are very close together, we have to choose a very small region around the highest peak, which leads to useless demixed signals. Most of the information of the desired signal is missing. Signal mixes which show a good separation in the histogram and fulfill the orthogonality condition on the time-frequency grid lead to better demixing results. For such signals we can achieve very good extracted signals, which are suitable for digital audio effects on just one source.

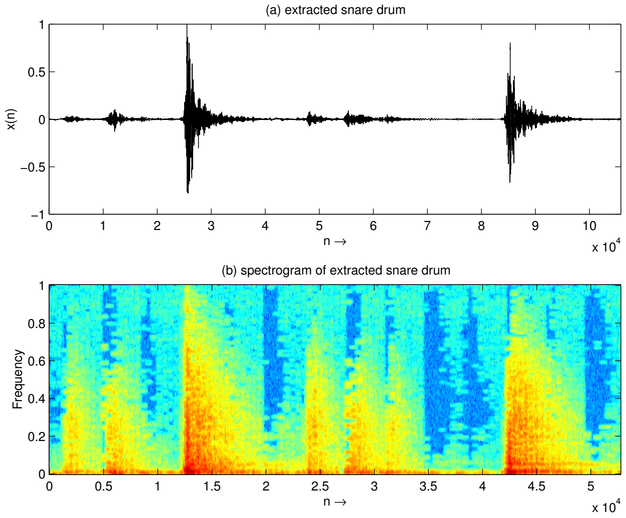

Figure 8: *Example for demixing a stereo drum signal.*

## 4. CONCLUSIONS

Source separation shows to be a useful pre-processing tool for the application of digital audio effects applied to different sources in a mix of many sources. The separation and extraction of up to four sources with acceptable quality is possible if a stereo microphone arrangement is used for recording. The transformed sources just have to mixed again by a combined amplitude and delay panning to achieve a mix with different effects on different sources. Several stereo mixes based on amplitude panning do not satisfy the orthogonality condition in time and frequency and are not suitable for source separation. Another approach by applying digital audio effects to specific areas of the time-frequency grid belonging to different sources without demixing is in the focus of current research.

## 5. REFERENCES

[1] J. Blauert, *Spatial Hearing*, MIT Press, Cambridge Mass., revised edition, 1996.

[2] J. Anemüller and T. Gramß, “Blinde akustische Quellentrennung im Frequenzbereich,” in *Fortschritte der Akustik - DAGA ’98*, Oldenburg, Germany, pp. 350–351, 1998.

[3] J. Anemüller, *Convolutive Blind Source Separation*, Ph.D. thesis, University of Oldenburg, Germany, 2001.

[4] A. Jourjine, S. Rickard, and Ö. Yilmaz, “Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures,” in *Proc. ICASSP-2000*, Istanbul, Turkey, Vol. 5, pp. 2985-88, 2000.

[5] S. Rickard and Ö. Yilmaz, “On the Approximate W-Disjoint Orthogonality of Speech,” in *Proc. ICASSP-2002*, Orlando, USA, Vol. 1, pp. 529–532, 2002.

[6] F. Keiler D. Arfib and U. Zölzer, “Time-frequency Processing,” in *DAFX–Digital Audio Effects*, U. Zölzer, Ed., pp. 237–263, J. Wiley & Sons, 2002.