Audio Segmentation Using Different Time-Frequency Representations

C. MEYER, M. SPIERTZ

Institute of Communications Engineering, RWTH Aachen University, 52056 Aachen, Germany

meyer@ient.rwth-aachen.de

Abstract. *A blind audio source separation technique for mono mixtures, composed of two noise-free instantaneously mixed audio sources, is presented. The separation is done by image segmentation of different time-frequency representations. Therefore a mixture is transformed by the Short-Time Fourier Transform into time-frequency domain. The resulting spectrogram image is used in log-amplitude or normal amplitude form and then segmented into significant areas. This areas are separately retransformed into time domain, and then grouped to segregate the source signals. The output results are rated by the improvement of SNR compared to the mixture.*

Keywords

Blind Source Separation, Spectrogram, Segmentation, Short-Time Fourier Transform, Peak Detection

1. Introduction

Normally, people are surrounded by simultaneous perceived sounds from different sources. They are able to focus their attention on a specific sound source e.g. an instrument in a concert or a voice on a party (*cocktail-party effect*). Great effort has been made to solve this problem with a computer, because there are a large number of possible applications in analysing, editing and manipulating audio data [4, 7, 8]. Examples are: Remastering of audio recordings, signal-preprocessing or -quality improvement e.g. for speech recognition or acoustic quality inspection of motors in noisy environments, intelligent hearing aids or automatic music transcription. In *Blind Source Separation* (BSS) only the mixture of the signals is available, and nothing is known about the sources or the mixing process like the room transfer function or the position of the sources.
Motivated by Virtanen [8] we segregate the sources by separating the time-frequency representation (spectrogram) of the mixture via image segmentation algorithms. The paper compares different ways to extract significant spectrogram-areas by adapting the spectrogram. The separated areas represent part-signals of different sources. They can afterwards be grouped to segregate the sources.
Instead of investigating the real-world case of a *convolutive mixture*, where mixtures of different sources are recorded in one room and each source contributes via multiple paths with different time delays and weights through reflexions and reverberations, we focus on the *instantaneous mixing* case. There the sources are mixed synthetically without any room characteristics. Hence no reflections are considered, so that the frequency-domain is more sparse.
Section 2 gives a short introduction to source separation, explains the separation via spectrogram and shows different peak detection methods to calculate positions of harmonics. In section 3 the experimental results are presented. Finally a conclusion is given in section 4.

2. Source Separation

2.1. Basics

Figure 1 shows the general scheme for source separation. The mixtures x_m, of N source signals s_n, are observed with M microphones. The mixtures are splitted into sub-signals p. In the case of oversegmentation the sub-signals are grouped in an appropriate way, for optimal outputs u_n.

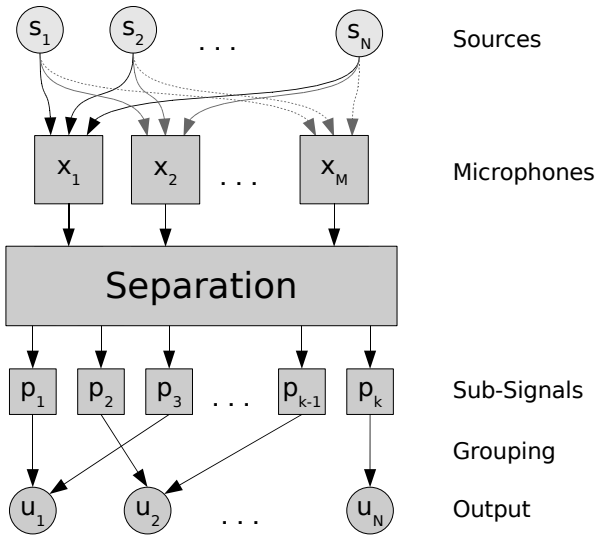

Fig. 1. General scheme of source separation.

The audio segmentation technique presented in this paper is an algorithm for mono mixtures composed of two sources (N=2, M=1). The source separation problem can generally be written as:

$$x(t) = A \cdot s(t) + n \tag{1}$$

Where x represents the observed mixed signal(s) as a vector. A is the unknown mixing matrix, $s = [s_1, s_2, ...s_N]^T$ the unknown source vector and n an additive noise term. Often the noise term is omited and the easier case of a noise-free environment is assumed – also in this paper.

Most common source separation methods use the concept of *Independent Component Analysis* (ICA), where normally more or equal sensors than sources are given to observe the mixed signals (for example with a microphone-array). When the source signals are statistically independent and non-Gaussian (except one source), ICA is able to separate the sources into statistically independent output signals by estimating a demixing matrix.

Beside this and other multiple input algorithms, one channel algorithms have been developed. They only need the input of a mono mixture x. An example is the *Independent Subspace Analysis* (ISA) [1] witch uses the ICA for decomposing a time-frequency representation of the signal into independent sub-components and combines the extracted parts. Other techniques (for one-channel source separation) use nonnegative matrix factorization or sparse coding [8]. We separate the sources by spectrogram segmentation as described in the next subsection.

2.2. Spectrogram Segmentation

Most parts of music comprise harmonic structures. The aim is to find and separate this harmonics plus other significant areas in a convenient time-frequency representation. Figure 2 illustrates a part of a spectrogram containing two time overlapping tones with different fundamental frequencies and different onsets. Harmonics are represented by horizontal lines (called tracks). A possible segmentation is shown in light gray. Here the frequency-borders are placed in the middle between two harmonics, and the time-borders are placed a few frames before the beginning and after the end of each harmonic. The benefit compared with other separation methods is, that the temporal structure of the source signals is automatically taken into account, because significant parts are cut out of the time-frequency domain.

A *Short-Time Fourier Transform* (STFT) is currently used to calculate the spectrogram. With the discrete circular frequency $\omega_k = 2\pi k/L$ the discrete STFT is:

$$\underline{X}(\omega_k, n) = \sum_{n=0}^{L-1} w(n)x(n-k)e^{-j\omega_k n} \tag{2}$$

A Hann window w is used to cut out a part of the time-signal x. The windows overlap with the half window length. The signal parts are each transformend via FFT into the frequency domain and each result depicts a column in the spectrogram. The absolute value of the complex-valued spectrogram $X = |\underline{X}|$ is taken to handle the image segmentation in the next step. The image is segmented by finding the harmonics via peak detection, and then using this positions (binary image: peak/no peak) to calculate the segment borders. After the segmentation process the borders of the found segments are used to split the original complex-valued spectrogram into parts. This parts are separately retransformed into the time domain, and can there (in the case of oversegmentation) be grouped to approximate the source signals. In the retransforming step the complex-valued spectrogram is used in order to obtain the original phase for each part. In the next subsection different peak detection methods, used to calculate the positions of the harmonics, are presented.

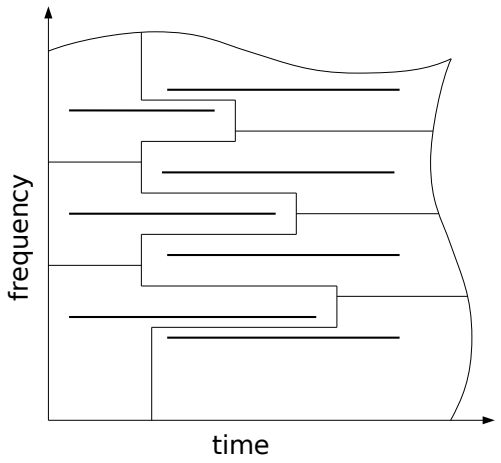

Fig. 2. Part of a spectrogram segmentation. The harmonics (horizontal lines; called tracks) of two single tones overlap in time. A possible segmentation is shown in light gray.

2.3. Peak Detection/Selection

The peak detection methods mentioned above are preceded by a cross-correlation function (CCF) in order to detect the sinusoidal partials (harmonics) [5]. The multiplication with the Hann window w in the time domain (involved with the STFT) corresponds to a convolution of its Fourier transform W in the frequency domain:

$$w(t) \cdot x(t) \circ\!\!-\!\!\bullet\ W(\omega) * X(\omega) \tag{3}$$

So cross-correlating each column in the spectrogram with W is equivalent to searching the harmonics (delta impulses):

$$\Gamma_{XW}(m) = \sum_n X(n)W(n+m) \tag{4}$$

Each maximum of the cross-correlation represents a possible position of a harmonic (sinusoidal partial), but it also could result from noise or attacks[1] in the signal.

[1] fast energy rise; transient oscillation

Strategies

Now three different strategies A), B) and C) to choose the right peaks are presented. They all have in common, that a balanced number of area-segments have to be found in the spectrogram. On the one hand a small number of segments leads to probably low separation quality and on the other hand a extremely high number of segments is difficult to arrange in the grouping step. Hence, found peaks which belong to the same harmonic have to be merged in order to result in one segment.

In order to find peaks, each time slot in Γ_{XW} (column in the preprocessed spectrogram) is observed separately. Searching for interesting peaks in each column implies not to detect every local maximum as a peak – so ignore noise and detect peaks corresponding with harmonics. Based on a peak detection algorithm from Eli Billauer[2] noise (smaller peaks) is ignored and peaks are found, without smoothing (e.g. by a low-pass filer) and therewith changing the original signal. The method will be called *detectHighPeaks* in this paper. Two criteria have to be fulfilled electing a local maximum to be an interesting peak:

- A local maximum is a peak, when it is more distant to the adjacent minima to the left and right than a defined threshold T.
- A minimum is the lowest point between two maxima.

A) The first strategy concentrates on the harmonics with the main signal energy. Therefore *detectHighPeaks* is applied to the CCF Γ_{XW}. A stable criterion is to use a percentage of the maximal amplitude from the whole spectrogram for the threshold T – a good value is 5%. In the next step a median filter is applied on each row of the binary image, because there are two problems left: Holes in the tracks[3] and single peaks with no adjacent peak. The median filter closes holes in the tracks and removes single peaks based on noise. A median filter of length five ensures balanced results.

In figure 3a) the spectrogram of a mixed cello and piano tone is presented. The piano tone begins circa at 0.6s and the cello tone at 0s. The found peaks are shown in 4a).

A2): If T is calculated for each column separately, by taking a percentage of the maximum amplitude of this column, also harmonics are indicated at times with lower signal energy. But if there are time frames which are comprised of noise, many peaks appear where no harmonics are.

B) With the next strategy also tracks at higher frequencies are found, by firstly processing Γ_{XW} with the logarithm:

$$\Gamma_{log} = \log_{10}(\Gamma_{XW} + \varepsilon) + a \tag{5}$$

The variable ε is a small value to prevent a zero argument, and a is positive value which effects the smallest value of Γ_{log} to be zero. The result is shown in figure 3b). In the next step *detectHighPeaks* is applied with a threshold T calculated for each column of the spectrogram. The value T is chosen to 1% of the maximal amplitude.

Now many peaks resulting from noise areas are present. Firstly all peaks are cleared which are located where the spectrogram X energy is lower that the average human auditory threshold. It can be approximated [2][6, p.243] with:

$$\frac{L_A(f)}{dB} = +3.64\left(\frac{f}{1000Hz}\right)^{-0.8} + 10^{-3}\left(\frac{f}{1000Hz}\right)^{4} - 6.5 \cdot exp\left[-0.6\left(\frac{f}{1000Hz} - 3.3\right)^{2}\right] \tag{6}$$

The *sound pressure level* (SPL) could be calculated from the spectrogram if the time signal is scaled with $max|x(t)| = 1$:

$$P = P_N + 10 \cdot \log_{10}|\underline{X}|^2 \tag{7}$$

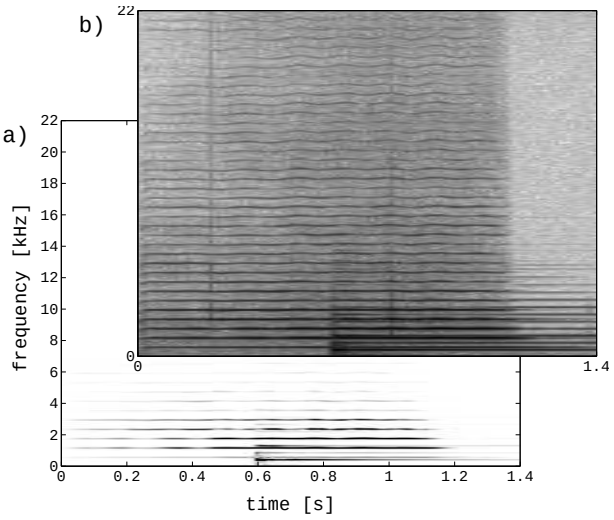

Fig. 3. Spectrogram of a mixed cello and piano tone. a) in normal amplitude form, b) in log-amplitude form. The harmonics can be recognized as lines.

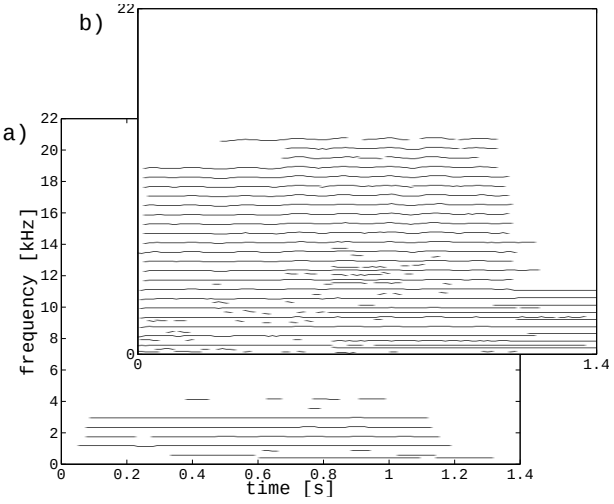

Fig. 4. Found peaks of the spectrogram from figure 3. a) with strategy A), b) with strategy B).

[2] http://www.billauer.co.il/peakdet.html

[3] Tracks are the lines in the spectrogram representing harmonics.

The term $P_N = 96dB$ is used to normalize the SPL like in audio CDs. So every peak where $P < L_A$ is cleared.
Secondly single peaks are cleared out which have no peak in there neighborhood. No median filter is used to solve this as described in A), because it would often break vibrato[4] tracks in high frequency (HF) regions. The reason is, that in a STFT vibrato is easier visible in the HF regions, because the percentage resolution is higher in this regions.
And finally the number of area-segments which would result from the current peaks is bounded to a limit B. Therefore for each area-segment the containing spectrogram energy (from X) is added up. This entries are sorted descending, and only peaks form the first B segments remain – and all others are cleared out. Now the area-segments are calculated from this remaining peaks, which are shown in figure 4b). Here we use a limit of $B = 100$ segments.

C) The third strategy uses the same peak-clearing techniques as describe in B), but the peaks are detected without applying the logarithm. Only *detectHighPeaks* is used with a very low threshold T, calculated for each spectrum separately. Here we use 0.01% of the maximum amplitude. The results difference only little from strategy B).

3. Experimental Results

The peak detection algorithms introduced in 2.3 are compared for instantaneous mixtures. Single cello and piano tone samples from *The University of Iowa Electronic Music Studios*[5] are converted to mono signals and then mixed pairwise. The 651 cello tones (339x arco + 312x pizzicato) and 516 piano tones are sampled with 44.1kHz. In the experiment 1000 random chosen pairs of a piano and a cello tone are mixed with five different time delays (0, 50, 100, 150, 200ms) between the two tone onsets. The instrument which starts with a delay is chosen randomly. Each mixture of these tones is transformed with the STFT via three different window-lengths (512, 1024 and 2048 samples). This is done for every peak detection method.
Each tone is normalized to identical maximal amplitude. The onsets are defined as the first time a signal passes over 10% of its maximal amplitude. The signals are limited to a maximal length of 3.4 seconds.
After the spectrogram separation the sub-signals must be grouped to segregate the original sources. In the moment we use the original source signals as references to group this sub-signals, in order to focus on the separating step and not on the grouping step. If more than 100 sub-signals result form the separation process the separation failed and the sub-signals are not grouped (compare last row in table 2).
To rate the separation quality quantitative, we use the *improvement signal-to-noise ratio* (ISNR) [3]. Therefore the SNR is calculated for each source, before and after the separation. The ISNR is the difference, and hence the improvement compared with the mixture:

$$\mathrm{SNR_{IN}} = 10 \cdot \log_{10}\left(\frac{\sum_t |s_j(t)|^2}{\sum_t |x(t) - s_j(t)|^2}\right) \tag{8}$$

$$\mathrm{SNR_{OUT}} = 10 \cdot \log_{10}\left(\frac{\sum_t |s_j(t)|^2}{\sum_t |u_j(t) - s_j(t)|^2}\right) \tag{9}$$

$$\mathrm{ISNR} = \mathrm{SNR_{OUT}} - \mathrm{SNR_{IN}} \tag{10}$$

The mean ISNR value is taken from both sources. The results for the explained strategies and different window-lengths are presented in figure 5. Because the highest window-lengths achieves the best results, this case is shown in figure 6 in detail. Table 1 shows the results for different delay times between the two tones, and in table 2 some information about the number of resulting sub-signals are listed.
Strategies which detect more harmonics reach better ISNR valus, than strategies which focus on main signal energy. On the other hand they also result in much more sub-signals which must be grouped. Here we used a reference grouping technique which is able to group even large number of segments. With a blind grouping technique this could look different, but the limit of 100 area-segments for strategy B) and C) is an good upper value for a blind grouping technique. Also it is pointed out, that higher window-lengths obtain better ISNR values. Table 1 shows, that the separation is more successful for longer delay-times between the two tones.

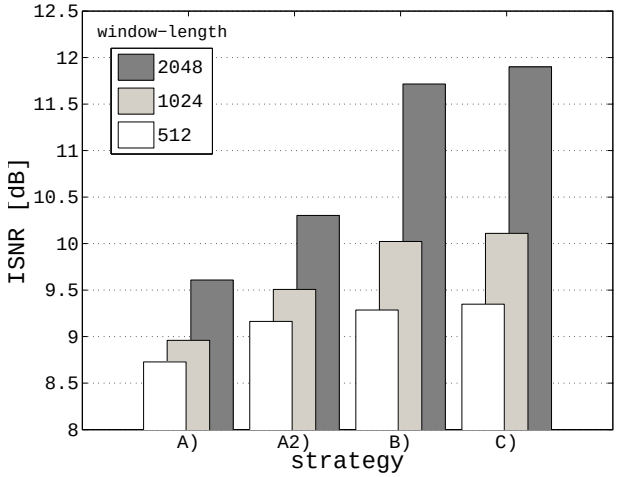

Fig. 5. Ratings for different strategies and window-lengths.

	A)	A2)	B)	C)
0 ms	8.3	8.7	9.3	9.4
50 ms	8.6	9.1	9.8	9.9
100 ms	9.1	9.7	10.4	10.5
150 ms	9.6	10.2	10.9	11.0
200 ms	10.0	10.6	11.3	11.4

Tab. 1. Results for different delay times in ISNR [dB].

[4] musical effect: tone pitch variation, like in the HF-parts in figure 3b).
[5] http://theremin.music.uiowa.edu

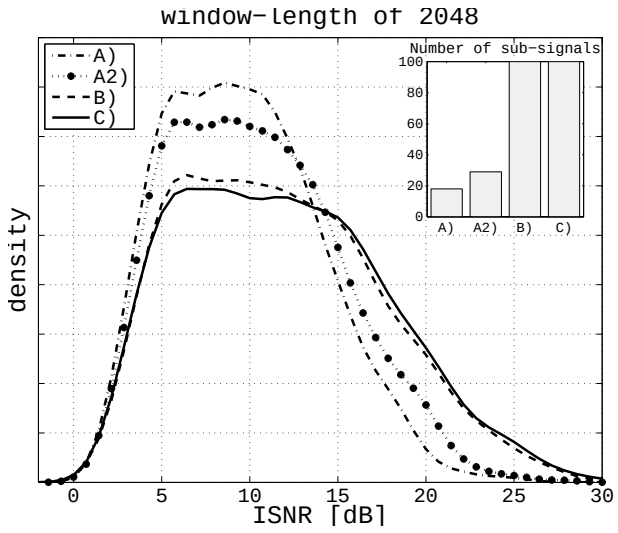

Fig. 6. Probability density of ISNR for different peak detection strategies and window-lengths of 2048.

	A)	A2)	B)	C)
min number of sub-sig.	1	2	95	77
mean number of sub-sig.	25	37	100	100
max number of sub-sig.	100	100	100	100
more than 100 sub-sig.	4.4%	10.4%	–	–

Tab. 2. Number of sub-signals of the different strategies.

4. Conclusion

A technique has been introduced which segregates mono mixtures via image segmentation of time-frequency representations. The spectrogram is divided into significant area-segments of different harmonics. The method gives the chance to uphold the temporal structure of the source signals, because the sub-signals represent areas in the time-frequency domain. Different strategies to detect harmonic positions have been shown and compared.

Different ways to clear out peaks of minor area-segments have been presented. They can bound the number of sub-signals. In the future more a priori knowledge to find important tracks could be used. For example the equidistant space between harmonics for one fundamental frequency. Or peaks which belong together could be grouped by using statistical dependencies.

The main challenge are the time-frequency regions with overpapping sources. In the moment the spectrogram is parted with hard borders. In the future we want to solve the overlapping problem with higher spectrogram resolutions, or by estimating the correct amplitudes in this regions.

Acknowledgements

The research described in the paper was supervised by Prof. J. R. Ohm, IENT, RWTH Aachen University.
The author wish to thank the team of the IENT, RWTH Aachen University for assistance and help.

References

[1] Casey, M., and Westner, A. Separation of mixed audio sources by independent subspace analysis.

[2] Guddeti, R. R., and Mulgrew, B. Perceptually Motivated Blind Source Separation of Convolutive Mixtures.

[3] Molla, M. I., and Hirose, K. Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum. 893–900.

[4] Pedersen, M., Larsen, J., Kjems, U., and Parra, L. A Survey of Convolutive Blind Source Separation Methods, 2007.

[5] Rodet, X. Musical sound signals analysis/synthesis: Sinusoidal+ residual and elementary waveform models.

[6] Terhardt, E. *Akustische Kommunikation. Grundlagen mit Hörbeispielen*. Springer Verlag, Berlin, Heidelberg, New York, 1998.

[7] Torkkola, K. Blind separation for audio signals – are we there yet. *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'99)* (1999), 239–244.

[8] Virtanen, T. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. 1066–1074.

About Authors...

C. MEYER was born in 1979 and studies electrical engineering at RWTH Aachen University. His current activity is writing his diploma thesis at the Chair and Institute of Communications Engineering (IENT).

M. SPIERTZ was born in 1980. He received his diploma degree in electrical engineering from RWTH Aachen University. Currently he is working towards his Ph.D. degree.