BLUES from Music: BLind Underdetermined Extraction of Sources from Music*

Michael Syskind Pedersen, Tue Lehn-Schiøler, and Jan Larsen

Intelligent Signal Processing, IMM, Technical University of Denmark
{msp, tls, jl}@imm.dtu.dk

Abstract. In this paper we propose to use an instantaneous ICA method (BLUES) to separate the instruments in a real music stereo recording. We combine two strong separation techniques to segregate instruments from a mixture: ICA and binary time-frequency masking. By combining the methods, we are able to make use of the fact that the sources are differently distributed in both space, time and frequency. Our method is able to segregate an arbitrary number of instruments and the segregated sources are maintained as stereo signals. We have evaluated our method on real stereo recordings, and we can segregate instruments which are spatially different from other instruments.

1 Introduction

Finding and separating the individual instruments from a song is of interest to the music community. Among the possible applications is a system where e.g. the guitar is removed from a song. The guitar can then be heard by a person trying to learn how to play. At a later stage the student can play the guitar track with the original recording. Also when transcribing music to get the written note sheets it is a great benefit to have the individual instruments in separate channels. Transcription can be of value both for musicians and for people wishing to compare (search in) music. On a less ambitious level identifying the instruments and finding the identity of the vocalist may aid in classifying the music and again make search in music possible. For all these applications, separation of music into its basic components is interesting. We find that the most important application of music separation is as a preprocessing step.

Examples can be found where music consists of a single instrument only, and much of the literature on signal processing of music deals with these examples. However, in the vast majority of music several instruments are played together, each instrument has its own unique sound and it is these sounds in unison that produce the final piece. Some of the instruments are playing at a high pitch and

* This work is supported by the Danish Technical Research Council (STVF), through the framework project "Intelligent Sound", STVF no. 26-04-0092, the PASCAL network, contract no. 506778. and the Oticon Foundation.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 392–399, 2006.

some at a low, some with many overtones some with few, some with sharp onset and so on. The individual instruments furthermore each play their own part in the final piece. Sometimes the instruments are played together and sometimes they are played alone. Common for all music is that the instruments are not all playing at the same time. This means that the instruments to some extent are separated in time and frequency. In most modern productions the instruments are recorded separately in a controlled studio environment. Afterwards the different sources are mixed into a stereo signal. The mixing typically puts the most important signal in the center of the sound picture hence often the vocal part is located here perhaps along with some of the drums. The other instruments are placed spatially away from the center. The information gained from the fact that the instruments are distributed in both space, frequency and time can be used to separate them.

Independent component analysis (ICA) is a well-known technique to separate mixtures consisting of several signals into independent components [1]. The most simple ICA model is the instantaneous ICA model. Here the vector $\mathbf{x}(n)$ of recorded signals at the discrete time index n is assumed to be a linear superposition of each of the sources $\mathbf{s}(n)$ as

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

where $\mathbf{A}$ is the mixing matrix and $\boldsymbol{\nu}(n)$ is additional noise. If reverberations and delays between the microphones are taken into account, each recording is a mixture of different filtered versions of the source signals. This model is termed the *convolutive* mixing model.

The separation of music pieces by ICA and similar methods has so far not received much attention. In the first attempts ICA was applied to separation of mixed audio sources [2]. A standard (non-convolutive) ICA algorithm is applied to the time-frequency distribution (spectrogram) of different music pieces. The resulting model has a large number of basis functions and corresponding source signals. Many of these arise from the same signal and thus a postprocessing step tries to cluster the components. The system is evaluated by listening tests by the author and by displaying the separated waveforms. Plumbley et al. [3] presents a range of methods for music separation, among these are an ICA approach. Their objective is to transcribe a polyphonic single instrument piece. The convolutive ICA model is trained on a midi synthesized piece of piano music. Mostly, only a single note is played making it possible for the model to identify the notes as a basis. The evaluation by comparing the transcription to the original note sheets showed good although not perfect performance. Smaragdis et al. has presented both an ICA approach [4] and a Non-negative Matrix Factorization (NMF) approach [5] to music separation. The NMF works on the power spectrogram assuming that the sources are additive. In [6] the idea is extended to use convolutive NMF. The NMF approach is also pursued in [7] where an artificial mixture of a flute and piano is separated and in [8] where the drums are separated from polyphonic music. In [9] ICA/NMF is used along with a vocal discriminant to extract the vocal.

Time-Frequency (T-F) masking is another method used to segregate sounds from a mixture (see e.g. [10]). In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years. Here, source separation is based on organizational cues from auditory scene analysis [11]. When the source signals do not overlap in the time-frequency domain, high-quality reconstruction can be obtained [12]. However, when there are overlaps between the source signals good separation can still be obtained by applying a binary time-frequency mask to the mixture [12, 13]. Binary masking is also consistent with constraints from auditory scene analysis such as people's ability to hear and segregate sounds [14]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [15]. T-F masking is applicable to source separation/ segregation using one microphone [10, 16] or more than one microphone [12, 13]. In order to segregate stereo music into independent components, we propose a method to combine ICA with T-F masking in order to iterative separate music into spatially independent components. ICA and T-F masking has previously been combined. In [17], ICA has been applied to separate two signals from two mixtures. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio.

Section 2 provides a review of ICA on stereo signals. In section 3 it is described how to combine ICA with masking in the time frequency domain. In section 4 the algorithm is tested on real music. The result is evaluated by comparing the separated signals to the true recordings given by the master tape containing the individual instruments.

2 ICA on Stereo Signals

In stereo music, different music sources (song and instruments) are mixed so that the sources are located at spatially different positions. Often the sounds are recorded separately and mixed afterwards. A simple way to create a stereo mixture is to select different amplitudes for the two signals in the mixture. Therefore, we assume that the stereo mixture $\mathbf{x}$ at the discrete time index n can be modeled as an instantaneous mixture as in eqn. (1), i.e.

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1N} \\ a_{21} \cdots a_{2N} \end{bmatrix} \begin{bmatrix} s_1(n) \\ \vdots \\ s_N(n) \end{bmatrix} + \begin{bmatrix} \nu_1(n) \\ \nu_2(n) \end{bmatrix}. \tag{2}$$

Each row in the mixing matrix $[a_{1i}\ a_{2i}]^T$ contains the gain of the i'th source in the stereo channels. The additional noise could e.g. be music signals which do not origin from a certain direction. If the gain ratio a_{1i}/a_{2i} of the i'th source is different from the gain ratio from any other source, we can segregate this source from the mixture. A piece of music often consists of several instruments as well as singing voice. Therefore, it is likely that the number of sources is greater than two. Hereby we have an *underdetermined* mixture. In [18] it was shown

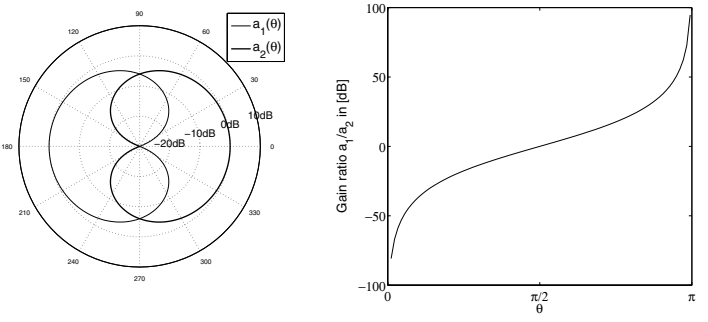

Fig. 1. The two stereo responses $a_1(\theta)$ and $a_2(\theta)$ are shown as function of the direction θ. The monotonic gain ratio is shown as function of the direction θ.

how to extract speech signals iteratively from an underdetermined instantaneous mixture of speech signals. In [18] it was assumed that a particular gain ratio a_{1i}/a_{2i} corresponded to a particular spatial source location. An example of such a location-dependant gain ratio is shown in Fig 1. This gain ratio is obtained by selecting the two gains as $a_1(\theta) = 0.5\big(1 - \cos(\theta)\big)$ and $a_2(\theta) = 0.5\big(1 + \cos(\theta)\big)$.

2.1 ICA Solution as an Adaptive Beamformer

When there are no more sources than sensors, an estimate $\tilde{\boldsymbol{s}}(n)$ of the original sources can be found by applying a (pseudo) inverse linear system, to eqn. (1).

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) = \mathbf{W}\mathbf{A}\mathbf{s}(n) \tag{3}$$

where $\mathbf{W}$ is a 2×2 separation matrix. From eqn. (3) we see that the output $\mathbf{y}$ is a function of $\mathbf{s}$ multiplied by $\mathbf{WA}$. Hereby we see that $\mathbf{y}$ is just a different weighting of $\mathbf{s}$ than $\mathbf{x}$ is. If the number of sources is greater than the number of mixtures, not all the sources can be segregated. Instead, an ICA algorithm will estimate $\mathbf{y}$ as two subsets of the mixtures which are as independent as possible, and these subsets are weighted functions of $\mathbf{s}$. The ICA solution can be regarded as an adaptive beamformer which in the case of underdetermined mixtures places the zero gain directions towards different groups of sources. By comparing the two outputs, two binary masks can be found in the T-F domain. Each mask is able to remove the group of sources towards which one of the ICA solutions places a zero gain direction.

3 Extraction with ICA and Binary Masking

A flowchart of the algorithm is presented in Fig. 2. As described in the previous section, a two-input-two-output ICA algorithm is applied to the input mixtures, disregarding the number of source signals that actually exist in the mixture. As shown below the binary mask is estimated by comparing the amplitudes of the two ICA outputs and hence it is necessary to deal with the arbitrary scaling

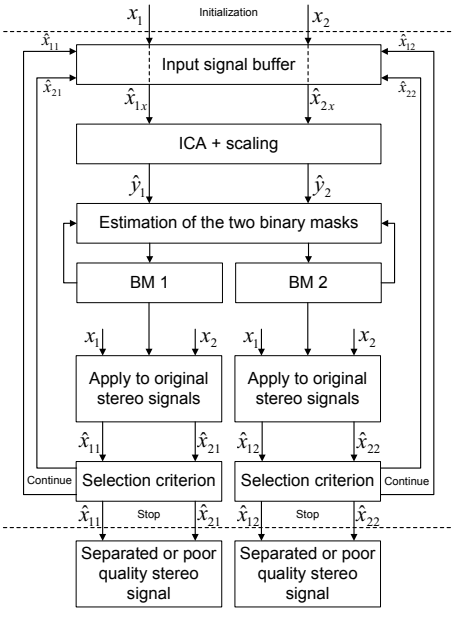

Fig. 2. Flowchart showing the main steps of the algorithm. From the output of the ICA algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the ICA step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal. The flowchart has been adopted from [18].

of the ICA algorithm. As proposed in [1], we assume that all source signals have the same variance and the outputs are therefore scaled to have the same variance. From the two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$, spectrograms are obtained by use of the Short-Time Fourier Transform (STFT):

$$y_1 \rightarrow \hat{y}_1 \rightarrow Y_1(\omega, t) \tag{4}$$

$$y_2 \rightarrow \hat{y}_2 \rightarrow Y_2(\omega, t), \tag{5}$$

where ω is the frequency and t is the time index. The binary masks are then found by a bitwise amplitude comparison between the two spectrograms:

$$\mathrm{BM1}(\omega, t) = \tau|Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{6}$$

$$\mathrm{BM2}(\omega, t) = \tau|Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{7}$$

where τ is a threshold that determines the sparseness of the mask. As τ is increased, the mask is sparser. We have chosen $\tau = 1.5$. Next, each of the two binary masks is applied to the original mixtures x_1 and x_2 in the T-F domain, and by this non-linear processing, some of the music signal are *removed* by one of the masks while other parts of music are removed by the other mask. After

the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT and and two sets of masked output signals $(\hat{x}_{11}, \hat{x}_{21})$ and $(\hat{x}_{12}, \hat{x}_{22})$ are obtained.

In the next step, it is considered whether the masked output signals consists of more than one signal. The masked output signals are divided into three group defined by the selection criterion in section 3.1. It is decided whether there is one signal in the segregated output signal, more than one signal in the segregated output, or if the segregated signal contains too little energy, so that the signal is expected to be of too poor quality.

There is no guarantee that two different outputs are not different parts of the same separated source signal. By considering the correlation between the segregated signals in the time domain, it is decided whether two outputs contains the same signal. If so, their corresponding two masks are merged. Also the correlation between the segregated signals and the signals with too poor quality is considered. From the correlation coefficient, it is decided whether the mask of the segregated signal is extended by merging the mask of the signal of poor quality. Hereby the overall quality of the new mask is higher.

When no more signal consist of more than one signal, the separation procedure stops. After the correlation between the output signals have been found, some masks still have not been assigned to any of the source signal estimates. All these masks are then combined in order to create a *background mask*. The background mask is then applied to the original two mixtures, and possible sounds that remain in the background mask are found. The separation procedure is then applied to the remaining signal to ensure that there is no further signal hidden. This procedure is continued until the remaining mask does not change any more. Note that the final output signals are maintained as stereo signals.

3.1 Selection Criterion

It is important to decide whether the algorithm should stop or whether the processing should proceed. The algorithm should stop separating when the signal consists of only one source or when the mask is too sparse so that the quality of the resulting signal is unsatisfactory. Otherwise, the separation procedure should proceed. We consider the covariance matrix between the output signals to which the binary mask has been applied, i.e. $\mathbf{R}_{xx} = \langle \mathbf{x}\mathbf{x}^H \rangle$. If the covariance matrix is close to singular, it indicates that there is only one source signal. To measure the singularity, we find the condition number of $\mathbf{R}_{xx}$. If the condition number is below a threshold, it is decided that $\mathbf{x}$ contains more than one signal and the separation procedure continues. Otherwise, it is assumed that $\mathbf{x}$ consists of a single source and the separation procedure stops.

4 Results

The method has been applied to different pieces of music. The used window length was 512, the FFT length was 2048. The overlap between time frames was 75%. The sampling frequency is 10 kHz. Listening tests confirm that the

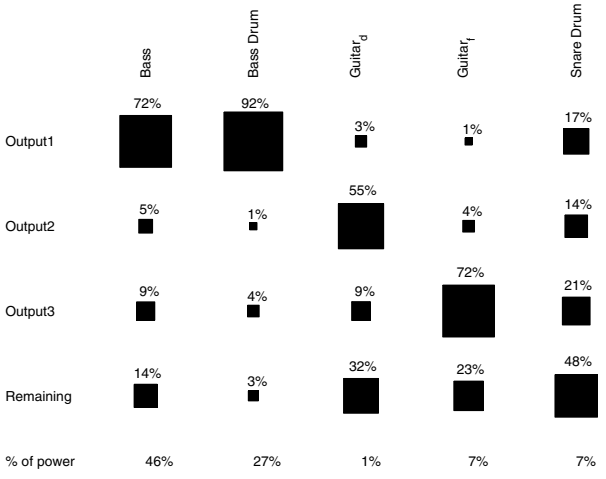

Fig. 3. Correlation coefficients between the extracted channels and the original stereo channels. The coefficients has been normalized such that the columns sum to one. The last row shows the percentage of power of the tracks in the mixture.

method is able to segregate individual instruments from the stereo mixture. We do not observe that correlations can be heard. However, musical artifacts are audible. Examples are available on-line for subjective evaluation [19]. In order to evaluate the method objectively, the method has been applied to 5 seconds of stereo music, where each of the different instruments has been recorded separately, processed from a mono signal into a stereo signal, and then mixed. We evaluate the performance by calculating the correlation between the segregated channels and the original tracks. The results are shown in Fig. 3 As it can be seen from the figure, the correlation between the estimated channels and the original channels is quite high. The best segregation has been obtained for those channels, where the two channels are made different by a gain difference. Among those channels is the guitars, which are well segregated from the mixture. The more omnidirectional (same gain from all directions) stereo channels cannot be segregated by our method. However, those channels are mainly captured in the remaining signal, which contains what is left when the other sources has been segregated. Some of the tracks have the same gain difference. Therefore, it is hard to segregate the 'bass' from the 'bass drum'.

5 Conclusion

We have presented an approach to segregate single sound tracks from a stereo mixture of different tracks while keeping the extracted signals as stereo signals. The method utilizes that music is sparse in the time, space and frequency domain by combining ICA and binary time-frequency masking. It is designed to separate tracks from mixtures where the stereo effect is based on a gain difference. Experiments verify that real music can be separated by this algorithm

and results on an artificial mixture reveals that the separated channel is highly correlated with the original recordings.

We believe that this algorithm can be a useful preprocessing tool for annotation of music or for detecting instrumentation.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
2. Casey, M., Westner, A.: Separation of mixed audio sources by independent subspace analysis. In: Proc. ICMC. (2000)
3. Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G., Sandler, M.B.: Automatic music transcription and audio source separation. Cybernetics and Systems **33** (2002) 603–627
4. Smaragdis, P., Casey, M.: Audio/visual independent components. Proc. ICA'2003 (2003) 709–712
5. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. Proc. WASPAA 2003 (2003) 177–180
6. Smaragdis, P.: Non-negative matrix factor deconvolution; extraction of multiple sound sourses from monophonic inputs. Proc. ICA'2004 (2004) 494–499
7. Wang, B., Plumbley, M.D.: Musical audio stream separation by non-negative matrix factorization. In: Proc. DMRN Summer Conf. (2005)
8. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In: Proc. EUSIPCO'2005. (2005)
9. Vembu, S., Baumann, S.: Separation of vocals from polyphonic audio recordings. In: Proc. ISMIR2005. (2005) 337–344
10. Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks **10** (1999) 684–697
11. Bregman, A.S.: Auditory Scene Analysis. 2 edn. MIT Press (1990)
12. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing **52** (2004) 1830–1847
13. Roman, N., Wang, D.L., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Amer. **114** (2003) 2236–2252
14. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi, P., ed.: Speech Separation by Humans and Machines. Kluwer, Norwell, MA (2005) 181–197
15. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In: Proc. ICASSP. (2000) 2985–2988
16. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks **15** (2004) 1135–1150
17. Kolossa, D., Orglmeister, R.: Nonlinear postprocessing for blind speech separation. In: Proc. ICA'2004, Granada, Spain (2004) 832–839
18. Pedersen, M.S., Wang, D.L., Larsen, J., Kjems, U.: Overcomplete blind source separation by combining ICA and binary time-frequency masking. In: Proc. MLSP. (2005) 15–20
19. http://www.intelligentsound.org/demos/demos.htm.