4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

EXTRACTION OF DRUM TRACKS FROM POLYPHONIC MUSIC USING INDEPENDENT SUBSPACE ANALYSIS

Christian Uhle *Christian Dittmar* *Thomas Sporer*

Fraunhofer AEMT, Langewiesener Strasse 22, 98693 Ilmenau, Germany
Email: {uhle,dittmarc,spo}@emt.iis.fhg.de

ABSTRACT

The analysis and separation of audio signals into their original components is an important prerequisite to automatic transcription of music, extraction of metadata from audio data, and speaker separation in video conferencing.

In this paper, a method for the separation of drum tracks from polyphonic music is proposed. It consists of an Independent Component Analysis and a subsequent partitioning of the derived components into subspaces containing the percussive and harmonic sustained instruments.

With the proposed method, different samples of popular music have been analyzed. The results show sufficient separation of drum tracks and non-drum tracks for subsequent metadata extraction. Informal listening tests prove a moderate audio quality of the resulting audio signals.

1. INTRODUCTION

1.1. Motivation

Audio related applications, like automatic transcription of music, analysis with respect to the extraction of metadata, application of audio effects to single instruments in a mix, and speaker separation in videoconferences, take advantage of the ability to separate single streams without a priori knowledge about the composition of the mixture.

Humans possess the ability to focus their attention on a single sound source within a mixture of multiple sources (”cocktail party effect”). While many of the factors which help to segregate streams have been investigated in the past (e.g. spatial distances between the sources, differences in pitch and sound quality, visual cues such as lip reading), the current state of the art is far from being able to mimic the capabilities of the human auditory system in this respect.

1.2. Technical introduction

The estimation of underlying signals from observations of their linear mixtures using a minimum of a priori information is called Blind Source Separation (BSS) [1]. BSS has gained much interest in various fields, including biomedical engineering, communication and auditory scene analysis. One of the different approaches to BSS is called Independent Component Analysis (ICA) and was introduced in the early 80s [2]. ICA assumes that the individual source signal are mutually statistical independent. This property is exploited in the algorithmic identification of the latent sources.

The ICA model expresses the observation signal $\boldsymbol{x}$ as the product of a mixing matrix $\boldsymbol{A}$ and a vector of statistically independent signals $\boldsymbol{s}$,

$$\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s} \tag{1}$$

Here, $\boldsymbol{A}$ is a $k \times l$ (pseudo-)invertible matrix with orthogonal columns, $\boldsymbol{s}$ is a random vector with l source signals and $\boldsymbol{x}$ a k-dimensional vector of observations with $k \geq l$. Further assumptions state that each source signal $\boldsymbol{s}$ is characterized by a stationary zero-mean stochastic process and only one of them has a Gaussian distribution. In the literature, a noise term is often added in the model on the input side, which is omitted for the sake of simplicity here. The in the model enclosed assumption, that there are at least as many observable mixture signals as source signals, limit the application of ICA to real world problems, where often fewer sensors than sources are available. Another limitation results through assuming that the components are mutually statistically independent.

To overcome these restrictions, various extensions to the basic ICA were proposed in the past ([3], [4], [5]). In the presented paper, Independent Subspace Analysis (ISA) is applied for the separation of the sources. Here, the components are divided into independent subspaces whose components are not independent. To meet the $k \geq l$ assumption, the input signal is transformed into a time-frequency representation (spectrogram) which can be interpreted as a multi-channel representation of the signal.

1.3. Related work

First approaches to apply ISA of time-frequency representations for the separation of different audio sources were described in [3] and [6]. In [6], this was done as a preprocessing step for subsequent audio analysis. An alternative approach to the partitioning of mixed audio signals into subspaces that correspond to individual sources works in the time domain and was presented in [7].

2. SYSTEM DESCRIPTION

Figure 1 shows an overview of the proposed method. A description of the preprocessing and the calculation of the independent components is given in detail in Section 3.

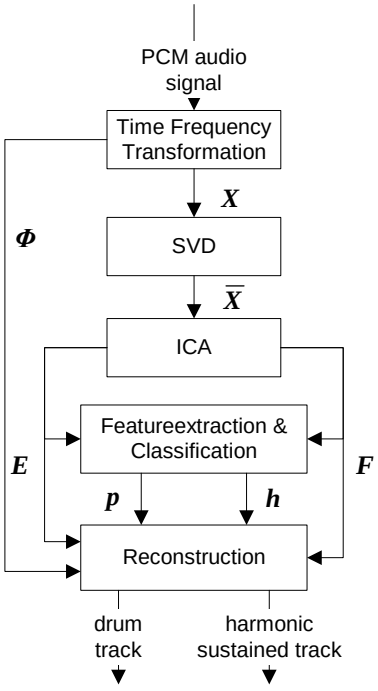

Figure 1: System overview

Section 4 will be dedicated to the description of the proposed novel approach to partitioning of the independent components into subspaces containing the percussive and harmonic sustained sounds.

3. DECOMPOSITION

3.1. Projection onto a maximally informative manifold

3.1.1. Pre-processing

The audio data is mapped to the spectral domain representation by a Short Time Fourier Transform (STFT) using a Hamming window [8]. From the computed complex valued spectrogram with n bins and m frames, the absolute values $\boldsymbol{X}$ and the phase information $\boldsymbol{\Phi}$ are derived. The phase information is omitted from the following calculations and is re-used only for re-synthesis at the end of the process.

3.1.2. Singular Value Decomposition

Consider the transposed spectrogram as the matrix $\boldsymbol{X}^T$, its Singular Value Decomposition (SVD) [9] is given by

$$\boldsymbol{X}^T = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{V}^T \tag{2}$$

The application of SVD is equivalent to the eigenvalue-decomposition of the covariance matrix of $\boldsymbol{X}^T$. Standard SVD algorithms return a diagonal matrix $\boldsymbol{D}$ of singular values in decreasing order and two orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}^T$. Matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, also referred to as the row basis, holds the left singular vectors, which equal the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$. Matrix $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, also referred to as the column basis, holds the right singular vectors equal to the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. The singular vectors are linearly independent and therefore provide the orthonormal basis for a rotational transform into the direction of the principal components.

3.1.3. Reduction of dimensionality

The SVD orders the basis vectors according to the size of their singular values. The singular values represent the standard deviations of the principal components of $\boldsymbol{X}$. These standard deviations are proportional to the amount of information contained in the corresponding principal components. A maximally informative subspace of the input data $\boldsymbol{X}$ is obtained by applying the following procedure.

A linear transformation matrix $\boldsymbol{T}$ is calculated according to equation (3), thereby $\overline{\boldsymbol{D}}$ is a submatrix consisting of the upper d rows of $\boldsymbol{D}$.

$$\boldsymbol{T} = \overline{\boldsymbol{D}} \cdot \boldsymbol{V}^T \tag{3}$$

The transformation matrix $\boldsymbol{T}$ is multiplied with the spectrogram $\boldsymbol{X}$, yielding a representation $\overline{\boldsymbol{X}}$ of reduced rank and maximally informative orientation (4).

$$\overline{\boldsymbol{X}} = \boldsymbol{T} \cdot \boldsymbol{X} \tag{4}$$

The number d of retained dimensions is a meaningful parameter of the whole process, which will be discussed later on. Observations show that a limited amount of 10 up to 30 dimensions is sufficient for the separation for analysis purpose. Fewer dimensions lead to an incomplete decomposition, while more dimensions give no reasonable improvement, increase the computational load and make the clustering process more complicated.

3.2. Independent Component Analysis

As stated earlier, the source separation model is a transformation, where the observations $\boldsymbol{x}$ are obtained by a multiplication of the source signals $\boldsymbol{s}$ by an unknown mixing matrix $\boldsymbol{A}$. The reduced rank spectrogram $\overline{\boldsymbol{X}}$ can be interpreted as an observation matrix, where each column is regarded as realizations of a single observation. In this work, the JadeICA algorithm [10] is applied for the estimation of $\boldsymbol{A}$. It minimizes higher order correlations by joint approximate diagonalization of eigenmatrices of cross cumulant tensors.

The estimated mixing matrix $\boldsymbol{A}$ is used to calculate the independent components. Its pseudo-inverse $\boldsymbol{A}^{-1}$ represents the unmixing matrix, by which the independent sources can be extracted. Employing equation (5) the independent temporal amplitude envelopes $\boldsymbol{E}$ are gained from the reduced rank spectrogram $\overline{\boldsymbol{X}}$.

$$\boldsymbol{E} = \boldsymbol{A}^{-1} \cdot \overline{\boldsymbol{X}} \tag{5}$$

The estimation of the independent frequency weights $\boldsymbol{F}$ is achieved by (6) and a subsequent pseudo-inversion.

$$\boldsymbol{F}^{-1} = \boldsymbol{A}^{-1} \cdot \boldsymbol{T} \tag{6}$$

The independent spectrograms are computed by multiplying one column of $\boldsymbol{F}$ with the corresponding row of $\boldsymbol{E}$,

$$\boldsymbol{S}_c = \boldsymbol{F}_{u,c} \cdot \boldsymbol{E}_{c,v} \tag{7}$$

where $u = 1,\ldots,n$, $v = 1,\ldots,m$ and $c = 1,\ldots,d$. The time signals of the separated sources are obtained by inverse STFT of the spectrograms $\boldsymbol{S}_c$.

4. GROUPING OF INDEPENDENT COMPONENTS INTO SUBSPACES

Depending on the total number of extracted components, the decomposition could be incomplete in the case of insufficient number of components or overcomplete in the contrary case. Incomplete representation leads to erroneous reconstruction $\tilde{\boldsymbol{x}}$ of the input signal from superposition of the entire set of independent components. Increasing the number of extracted components yields a smaller root-mean-squared error between the original audio signal $\boldsymbol{x}$ and the reconstruction $\tilde{\boldsymbol{x}}$ consisting of all components.

The perceptual impression of incomplete decomposition to a human listener is the disappearance of the weaker and sparser occurring sounds. On the other hand, an overcomplete decomposition causes the statistically most prominent instruments to occupy more extracted components than others. The problem arises to find a trade-off between loss of detail and extraction of quasi-redundant components. An overcomplete decomposition with subsequent partitioning of the components into distinct subspaces bypasses this drawback.

Regarding the aim of this work, classifying the components into two sets is appropriate. The features employed for this purpose are described in the following section.

4.1. Features

4.1.1. Percussiveness

A so-called percussiveness feature is extracted from the time-varying amplitude envelopes. Percussive components are assumed to feature impulses of fast attack and (slower) decay. Amplitude envelopes of components originating from harmonic sustained sounds feature plateaus. The percussive impulses are modelled using a simplistic model with an instantaneous "attack" and a linear decay towards zero within ca. 200 ms. This model template is convolved with the local maxima of the amplitude envelope.

The correlation coefficient of model and original vector represents the degree of percussiveness. Figures 2 and 3 show exemplary amplitude envelopes and model vectors of a percussive respectively harmonic sustained component.

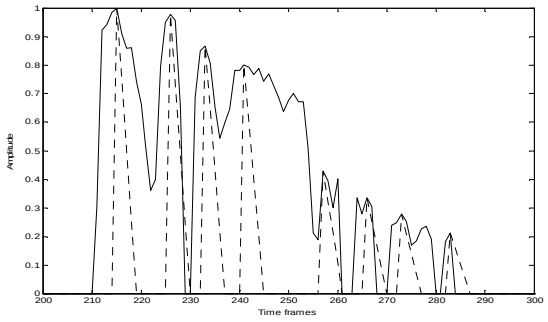

Figure 2: Amplitude envelope (solid) and model vector (dashed) of a harmonic sustained component

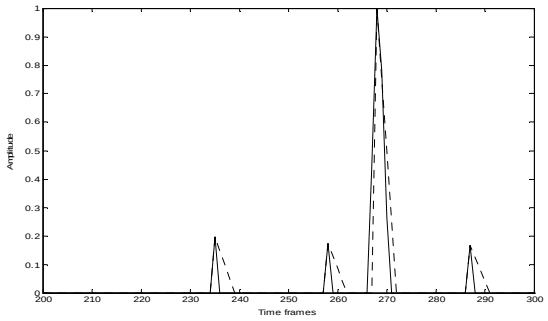

Figure 3: Amplitude envelope (solid) and model vector (dashed) of a percussive component

4.1.2. Noise-likeness

Noise-Likeness is derived from the frequency vector of the components. It is extracted to quantify the degree of noisiness which is assumed to indicate the affiliation of a component to the percussive subspace.

According to the findings of W. Aures [11] and R. Parncutt [12], (stating that tonal spectra feature narrow-band partials in contrast to atonal or noisy spectra), the following extraction method is proposed:

The local maxima of a frequency vector are convolved with a gaussian impulse with zero mean and variance σ, given by

$$g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{8}$$

Noise-likeness is estimated as the correlation coefficient of original and resulting model vector.

Figures 4 and 5 show a frequency vector and a model vector of a percussive and harmonic sustained components, respectively.

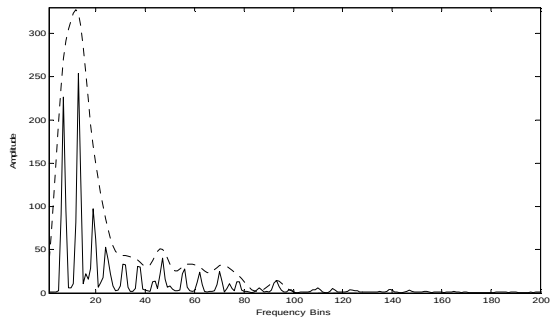

Figure 4: Frequency vector (solid) and model vector (dashed) of a harmonic sustained component

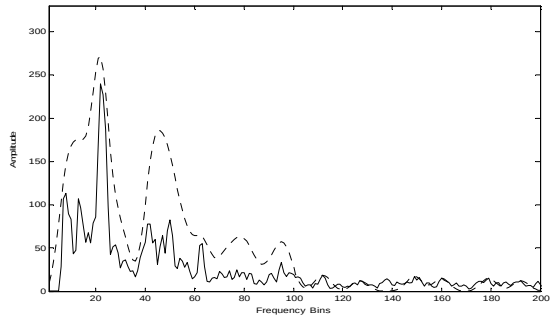

Figure 5: Frequency vector (solid) and model vector (dashed) of a percussive component

4.1.3. Spectral dissonance

Percussive sounds are supposed to contain more non-harmonic partials than harmonic sustained sounds. This assumption motivated the extraction of a dissonance measure according to Sethares [13]. It is derived from the frequency weights of independent components by adding the pairwise dissonance of all spectral components. The dissonance d of two sinusoids with frequencies f_1, f_2 and amplitudes a_1 respectively a_2 is given by

$$d(f_1, f_2, a_1, a_2) = a_1 a_2 \left(e^{-as(f_2-f_1)} - e^{bs(f_2-f_1)}\right) \tag{9}$$

with $a = 3.5$, $b = 5.75$ and $s = \frac{0.24}{(0.021 f_1 + 19)}$ as proposed in [13].

4.1.4. Spectral Flatness Measure

The Spectral Flatness Measure (SFM) describes the flatness properties of the spectrum of an audio signal. Percussive components are assumed to have flatter spectra than harmonic sustained components. Following the definition proposed in [14], SFM is calculated as the ratio of geometrical to arithmetical mean of the power spectrum coefficients (10).

$$SFM(x) = \frac{\sqrt[N]{\prod_{n=1}^{N} \boldsymbol{x}_n}}{\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n} \tag{10}$$

Thereby, *N* denotes the number of frequency bins of the power spectrum $\boldsymbol{x}$. The vector $\boldsymbol{x}$ equals one element-wise squared column vector of $\boldsymbol{F}$.

4.1.5. Third order cumulant

The use of higher order statistics on the independent components as a distance measure between the components is described in [7].

The independent frequency weights are transformed into the time domain using an inverse STFT. Subsequently, the third order cumulant is computed from the obtained time signal (11).

$$TOC(x) = E\{\boldsymbol{x}^3\} - 3E\{\boldsymbol{x}^2\}E\{\boldsymbol{x}\} + 2[E\{\boldsymbol{x}\}]^3 \tag{11}$$

$E\{\cdot\}$ denotes the expectation operator and $\boldsymbol{x}$ the time signal.

4.2. Classification

The set affiliations of the components are derived using a cascade of threshold-based decisions.

Starting with the most reliable feature, all components with feature values above a predefined threshold, which are therefore assumed to contribute to the percussive subspace, are passed to the next decision stage, while the others are rejected. The number of spuriously accepted components is allowed to be high. Further reduction is made by comparing additional features with their

respective threshold. The classification procedure results in two index vectors $\boldsymbol{p}$ and $\boldsymbol{h}$, where $\boldsymbol{p}$ contains the indices of the components classified as percussive and $\boldsymbol{h}$ all other components, assumed to originate from harmonic sustained sounds.

4.3. Stream Generation

Once the independent components are assigned to either the percussive or the harmonic sustained set, two independent audio streams are generated. This is achieved by a matrix multiplication of subsets of $\boldsymbol{E}$, containing the time varying amplitude envelopes, and $\boldsymbol{F}$, holding the static frequency weights, given by $\boldsymbol{p}$ and $\boldsymbol{h}$. The expression $\boldsymbol{S}_p = \boldsymbol{F}_{u,p}\boldsymbol{E}_{p,v}$ yields a reconstructed amplitude spectrogram corresponding to the percussive part of the input spectrogram $\boldsymbol{X}$ and accordingly $\boldsymbol{S}_h = \boldsymbol{F}_{u,h}\boldsymbol{E}_{h,v}$ represents the harmonic sustained part.

The matrix $\boldsymbol{\Phi}$, containing the phase information obtained before the ISA procedure, is element-wise multiplied in given by $\tilde{\boldsymbol{S}}_{u,v} = \boldsymbol{S}_{u,v}\left(\cos\left(\boldsymbol{\Phi}_{u,v}\right) + j\sin\left(\boldsymbol{\Phi}_{u,v}\right)\right)$. The re-synthesis to the time domain is achieved by windowed inverse STFT.

5. RESULTS

5.1. Test data

To evaluate the abilities of the presented approach, drum tracks were extracted from a database of 9 samples. The duration of each of these test items is 10 seconds. The test files were recorded at 44100 Hz sampling rate with 16 bits amplitude resolution. Different musical genres are represented within these examples, ranging from big band to rock/pop to electronica. They were chosen because of their quite different musical characteristics with respect to properties such as tempo, percussivity, instrumentation and melodicity.

To evaluate the performance of the method and for comparison of the investigated features, the extracted components were classified manually by listening to each resynthesized stream per test file in order to obtain the reference index vectors $\bar{p}$ and $\bar{h}$.

5.2. Comparison of the computational and manual classification of single features

To evaluate the performance of a single feature, two measures are given here. The quantity *found* is obtained according to equation (12) and represents the amount of correctly classified percussive components. To describe the misclassifications made, the quantity *added* is calculated as described in equation (13).

$$found = \frac{\sum size\left(p \cap \bar{p}\right)}{\sum size\left(\bar{p}\right)} \tag{12}$$

$$added = \frac{\sum size\left(p \cap \bar{h}\right)}{\sum size\left(\bar{h}\right)} \tag{13}$$

The following table shows the average of *found* and *added* values for the single features over all test files. Obviously, noise-likeness is the most reliable feature, while SFM is the weakest feature for this task.

Feature	*found* in %	*added* in %
Noise-likeness	95	71
Spectral dissonance	90	63
Percussiveness	80	48
Third order cumulant	80	45
SFM	70	59

Table 1: Performance of each of the single features (the threshold for each feature was optimized separately).

5.3. Comparison of the computational and manual classification of the feature combination

To quantify the reliability of the entire classification method, the same measures are applied as in the preceding section. With an appropriate choice of thresholds, a performance of 73% *found* and 26% *added* was measured on the described test data. In general, the separating capabilities are dependent on the nature of the input data. Less prominent sounds may appear to be proportionately decomposed into contrasting components.

Problems were also encountered in densely instrumented audio mixes, resulting in the absence of medium prominent percussive sounds and proportions of harmonic sustained sounds in the separated drum track.

6. CONCLUSIONS AND FUTURE WORK

In this paper a method for the extraction of drum tracks from polyphonic music was presented. For most of the test data, the separated tracks have sufficient audio quality and the amount of harmonic sustained sounds is tolerable for a subsequent extraction of metadata.

Improvements should be made in the classification stage, there a conventional classifier scheme, e.g. a nearest-neighbor classifier would be appropriate. Additional improvements are expected from the substitution of the STFT by a perceptually tuned filterbank for the time-frequency transform.

A preceding estimation of the “density” of the audio mix, i.e. the amount of occurring instruments, sounds and voices in the music, could lead to an appropriate choice of the number of estimated components.

The separation of the distinct percussive instruments into single streams will be one of our main concerns in the following research on that topic.

7. ACKNOWLEDGMENTS

The work on metadata extraction was co-funded by the Thüringen Ministery of Science, Research and Culture, and Thomson Multimedia. The authors wish to thank Juergen Herre for his proofreading of the paper and valuable suggestions to its clarity.

8. REFERENCES

[1] J. Karhunen, "Neural approaches to independent component analysis and source separation", *Proceedings of the European Symposium on Artificial Neural Networks*, pp. 249-266, Bruges, 1996.

[2] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, Wiley&Sons, 2001.

[3] M.A. Casey and A. Westner, "Separation of Mixed Audio Sources by Independent Subspace Analysis", *Proceedings of the International Computer Music Conference*, Berlin, 2000.

[4] J.-F. Cardoso, "Multidimensional independent component analysis", *Proceedings of ICASSP'98*, Seattle, 1998.

[5] A. Hyvärinen, P.O. Hoyer and M. Inki, "Topograghic Independent Component Analysis", *Neural Computation*, 13(7), pp.1525-1558. 2001.

[6] I.F.O. Orife, *Riddim: A rhythm analysis and decomposition tool based on independent subspace analysis*, Master thesis, Darthmouth College, 2001.

[7] S. Dubnov, "Extracting Sound Objects by Independent Subspace Analysis", *Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Helsinki, 2002.

[8] F.J. Harris, "On the use of windows for harmonic analysis with discrete Fourier Transform", *Proceedings of the IEEE*, vol. 66, no. 1, 1978.

[9] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1992.

[10] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for nonGaussian signals", *IEE Proceedings*, Vol. 140, no. 6, pp. 362-370, 1993.

[11] W. Aures, *Berechnungsverfahren für den Wohlklang beliebiger Schallereignisse, ein Beitrag zur gehörbezogenen Schallanalyse*, Dissertation (in german), Technical University Muenchen, 1984.

[12] R. Parncutt, *Harmony: A Psychoacoustical Approach*, Springer, NewYork, 1989.

[13] W. Sethares, "Local Consonance and the Relationship between Timbre and Scale", *J. Acoust. Soc. Am.*, 94 (3), pt. 1, 1993.

[14] International Standards Organization (ISO), *Information Technology – Multimedia Content Description Interface – Part 4: Audio*, ISO-IEC 15938-4 (E), 2001.