Monaural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization

Tuomas O. Virtanen, *tuomas.virtanen@tut.fi*
Tampere University of Technology, Institute of Signal Processing

Abstract — **A data-adaptive algorithm for the separation of sound sources from one-channel signals is presented. The algorithm applies weighted non-negative matrix factorization on the power spectrogram of the input signal. Perceptually motivated weights for each critical band in each frame are used to model the loudness perception of the human auditory system. The method compresses high-energy components, and enables the estimation of perceptually significant low-energy characteristics of sources. The power spectrogram is factorized into a sum of components which have a fixed magnitude spectrum with a time-varying gain. Each source consists of one or more components. The parameters of the components are estimated by minimizing the weighted divergence between the observed power spectrogram and the model, for which a weighted non-negative matrix factorization algorithm is proposed. Simulation experiments were carried out using generated mixtures of pitched musical instrument samples and percussive sounds. The performance of the proposed method was compared with other separation algorithms which are based on the same signal model. These include for example independent subspace analysis and sparse coding. According to the simulations the proposed method enables perceptually better separation quality than the existing algorithms. Demonstration signals are available at http://www.cs.tut.fi/~tuomasv/.**

Index Terms — **Sound source separation, non-negative matrix factorization, sparse coding, independent subspace analysis**

I. INTRODUCTION

In real-world audio signals several sound sources are usually mixed. The process in which individual sources are estimated from the mixture signal is called sound source separation. Separation of mixed sounds has several applications in the analysis, editing and manipulation of audio signals. These include for example object-based audio coding, automatic transcription of music, and computational auditory scene analysis. There are powerful algorithms for the processing of isolated sounds, therefore the capability of separating sources from polyphonic mixtures is very appealing. In this paper the focus is on the separation of music signals.

The definition of a sound source depends somewhat on the application. Usually the term is used to refer to an individual physical source or to an entity that humans perceive individually. Humans tend to perceive similar-souding physical sound sources as a single entity. For example, if a violin section plays in unisono, all the sounds arriving from the same musical instrument are perceived as a single source.

Humans are extremely skillful in "hearing out" individual sources from complex mixtures even in noisy conditions. Computational modeling of this ability is very difficult. All the existing separation systems are limited in either polyphony or quality. The most successful ones are those which try to extract only the most prominent source [1], [2].

Without any prior knowledge of the sources, the problem of estimating several overlapping sources from one input signal is ill-defined. By making some assumptions of the underlying sounds, it is possible to analyze and synthesize signals which are perceptually close to the original ones before mixing. For example the harmonicity of sources has been assumed in most systems which are aimed to separate musical sounds [1], [3].

Independent Component Analysis (ICA) has been successfully used to solve blind source separation problems in several application areas. A related technique called independent subspace analysis (ISA) has been used for sound separation e.g. by Casey and Westner [4], FitzGerald, Coyle, and Lawlor [5], Uhle, Dittmar, and Sporer [6], Orife [7], and with some modifications by Emmanual and Rodet [8]. The method tries to find source spectra which are statistically independent from each other. Also the analysis procedure proposed by Brown and Smaragdis [9] can be viewed as an ISA algorithm. A sound recognition system based on ISA has also been adopted in the MPEG-7 standardization framework [10].

The non-negative matrix factorization (NMF) algorithms proposed by Lee and Seung [11] has been suggested as a solution for the blind source separation problem with non-negativity constraints. The algorithm has been used for sound source separation by Smaragdis and Brown [12]. In sound classification NMF was reported to perform better than ISA [13].

A data-driven technique called sparse coding has been successfully used for example to model the functioning of the early stages of vision [14]. The term *sparse* is used to refer to a signal model, in which the data is represented in terms of a small number of active elements chosen out of a larger set. Sparse coding has been used for audio signal separation by Plumbley et al. [15], Abdallah [16], and Benaroya, Donagh,

The author is with the Institute of Signal Processing, Tampere University of Technology. This work was supported by the TISE graduate school and the Nokia Foundation.

Bimbot, and Gribonval [17].

The non-negative sparse coding algorithm proposed by Hoyer [18] combines non-negative matrix factorization and sparse coding. With some modifications and a temporal continuity objective the algorithm was used in sound separation by Virtanen [19]. More complex models derived from NMF and sparse coding have been proposed by Smaragdis [20] and Virtanen [21].

All the above mentioned studies claim that it is possible (to some degree) to separate sound sources without any prior knowledge of the sources, while it is clear that for robust high-quality separation more assumptions have to be made. In this paper, algorithms based on a linear signal model are discussed and compared. The proposed perceptually weighted NMF algorithm is shown to outperform previously proposed algorithms in perceptual separation quality.

A. Signal model

The signal model used in this paper and in ISA, NMF, and sparse coding methods is in general linear: each observation vector $\boldsymbol{x}_t$ is assumed to be a linear mixture of basis vectors $\boldsymbol{s}_n$. In the case of NMF and sparse coding the model is not necessarily noise-free. With a residual term $\boldsymbol{r}_t$ the model can be written as:

$$\boldsymbol{x}_t = \sum_{n=1}^{N} a_{t,n}\boldsymbol{s}_n + \boldsymbol{r}_t \qquad t = 1 \ldots K \tag{1}$$

where $a_{t,n}$ is the mixing weight of the n^{th} component in the t^{th} observation, N is the number of components, and K is the number of observations. In this paper, term *component* is used to refer to one basis function. A sound source is represented as a sum of one or more components.

In the signal model (1), only the observations are known. The algorithms for estimating the basis functions and weights are shortly reviewed in Section B. In a matrix form the model can be expressed as

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{R} \tag{2}$$

where $\boldsymbol{X} = \left[\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_K\right]^T$, $[\boldsymbol{A}]_{tn} = a_{t,n}$, $\boldsymbol{S} = \left[\boldsymbol{s}_1 \ \ldots \ \boldsymbol{s}_K\right]^T$, and $\boldsymbol{R} = \left[\boldsymbol{r}_1 \ \ldots \ \boldsymbol{r}_K\right]^T$. In this paper notation $[\boldsymbol{A}]_{tn}$ is used to refer the $(t,n)^{\text{th}}$ element of matrix $\boldsymbol{A}$.

In the case of music signals the most obvious choice for the observation matrix is a time-frequency spectrogram, so that the basis functions are the spectra of components. The components are parametrized by a fixed spectrum over time so that only the gains are time-varying. The signal model in (1) is very restrictive in the sense that natural sound sources have to be represented with a fixed spectrum over time. For example different fundamental frequencies of a source have different spectra. In practise this is solved by representing one sound source as a sum of several components. The observed data is separated into components, which are then clustered into sound sources, as suggested e.g. by Casey and Westner [4] and Virtanen [19]. In general the clustering is a difficult task and will not be discussed in this paper.

The time-domain signals and therefore also the complex spectra of sources sum linearly. However, the phase spectra of natural sounds are very unpredictable. Also, the human sound perception is rather insensitive to phase. The estimation of frame-wise phases for each source would make the model too complex, therefore complex spectra cannot be used in this framework. The linear addition of complex spectra has to be approximated as a linear addition of real-valued spectra.

When arbitrary complex sources with unknown random phase spectra are summed, the expectation value for the power spectrum of the sum equals the sum of the power spectra of the sources, if the phase spectra of the sources are independent. This holds only for the power spectra, not for magnitude spectra, for example. Thus, the most accurate linear addition is obtained in the power spectral domain.

Unfortunately, the power spectrogram is not a perceptually well motivated representation. The human sound perception is very nonlinear as a function of the intensity of the input signal, which enables the perception of low-intensity sources. The perceived loudness is approximately proportional to the logarithm of the intensity, or, to the intensity raised to the power 0.23 [28, pp. 181-214]. If the parameters are fitted to the observations in the power spectral domain, high-intensity observations will dominate the separation.

Basically the same estimation algorithms can be used for magnitude or power spectra. Taking element-wise an arbitrary power of the input spectrogram changes the summation properties and the dynamic range of the data. With this approach there is always a trade-off between the accurate summation in the power spectral domain and the sensitivity to low-intensity sources in the logarithmic domain, which can not be achieved at the same time. In most data-driven algorithms the model fitting is done using the magnitude spectra, which is a compromise between the alternatives.

B. Estimation criteria

In ISA, the estimation is done by assuming statistical independence of the spectra of components. The estimation can be done using any ICA algorithm, which usually consist of whitening, dimension reduction by principal component analysis, and estimation of an unmixing matrix which maximizes the independece of resultant spectra. The components are obtained by multiplying the observation matrix by the unmixing matrix. If the basis vectors are magnitude or power spectra, it is reasonable to restrict that the component matrix $\boldsymbol{S}$ and mixing weights $\boldsymbol{A}$ are element-wise nonnegative. Currently there do not exist algorithms which could estimate independent components which are restricted to be nonnegative and have non-negative mixing weights. This limitation has been tried to overcome by Plumbley and Oja [22], whose non-negative PCA algorithm can be use to estimate non-negative sources with real mixing weights. Even though the nonnegativity restrictions are not met with current ICA algorithms, the separation results of ISA are relatively good. In the simulation experiments presented in Section III, the proposed method is compared with

ISA where the independent components are estimated using FastICA [23], [24] and Jade [25], [26] algorithms.

In the NMF algorithms proposed by Lee and Seung [11], the non-negative basis functions and mixing weights are estimated using an iterative procedure which is based on the minimization of the Euclidean distance between the observed data $\boldsymbol{X}$ and model $\boldsymbol{AS}$, or divergenge D, given as

$$D(\boldsymbol{X}||\boldsymbol{AS}) = \sum_{t,f} d([\boldsymbol{X}]_{tf}, [\boldsymbol{AS}]_{tf}) \quad (3)$$

where the function d is defined as

$$d(p, q) = p\log\frac{p}{q} - p + q \quad (4)$$

The divergence is lower bounded by zero, which is attained only when $\boldsymbol{X} = \boldsymbol{AS}$. The divergence reduces to the Kullback-Leibler divergence when $\sum_{t,f}[\boldsymbol{X}]_{tf} = \sum_{t,f}[\boldsymbol{AS}]_{tf} = 1$ [11]. In this paper the use of the divergence is motivated by two factors: first, there exists an efficient algorithm for minimizing the divergence. Second, the divergence is less sensitive to large-energy observations than for example the Euclidean distance. This enables the use of power spectrogram as the observation.

Hoyer combined the multiplicative update used in the NMF with steepest descent [18]. His algorithm allows the use of further assumptions such as sparseness and temporal continuity of sources [19]. In sound source separation, the sparseness objective is used either for the mixing weights $a_{t,n}$ or for the spectra of components. For the mixing weights the sparseness objective means that the probability for mixing weight $a_{t,n}$ being zero is high. In matrix form (2) this means that mixing matrix $\boldsymbol{A}$ is sparse. In sound source separation the sparseness of sources means that only few sources are active at each time.

C. Improvements in the proposed method

As stated earlier, human sound perception is very nonlinear as a function of the intensity of sources, enabling the perception of low-energy sources. Thus, data-driven algorithms are not always able to extract low-energy sources even though they are perceptually significant. This phenomenon has been noticed for example by FitzGerald [27] in the case of ISA and percussice sound separation.

In this paper a method for modeling the loudness perception of the human auditory system is incorporated into the non-negative matrix factorization algorithm. The divergence is weighted linearly to obtain an error criterion which approximates the intensity perception of the human auditory system. The method is explained in Section II.A. It achieves the accurate summation in the power spectral domain and a good sensitivity on a large dynamic range at the same time.

The weighted non-negative matrix factorization algorithm for the estimation of the components is explained in Section II.B, and the synthesis procedure is shortly described in Section II.C. Especially the use of perceptually motivated weights increases the quality of the separated sources, as the simulation experiments presented in Section III indicate.

II. PROPOSED METHOD

An input signal is represented using the power spectrogram, which is calculated as follows. First, the time-domain signal is divided into frames and windowed. In our implementation a fixed 40 ms frame size is used with 50% overlap between frames since it provides a good compromise between time and frequency resolutions. The square root of the Hanning window is used, since it allows smooth synthesis using the overlap-add as explained in Section C.

Each frame is transformed into frequency domain by taking the discrete Fourier transform (DFT). The length of the DFT is equal to the frame size. Only positive frequencies are retained. Phases are discarded by squaring the absolutive values of the DFT spectra to result in the spectrogram $x_{f,t}$, where $f = 1...F$ is the discrete frequency index and $t = 1...K$ is the frame index. F is the number of frequency bins and K is the number of frames. Matrix $[\boldsymbol{X}]_{tf} \cong x_{t,f}$ is used to denote the observed power spectrogram. The phase spectrogram is stored since it is needed in the synthesis.

The source model used in the proposed method is given by Equations (1) and (2), so that the basis vectors are the power spectra of each component, and the mixing weights are the time-varying gains. The parameters are estimated using a weighted version of divergence as given by (3). The weighted divergence is defined as

$$D(\boldsymbol{X}||\boldsymbol{AS};\boldsymbol{W}) = D(\boldsymbol{W}.*\boldsymbol{X}||\boldsymbol{W}.*(\boldsymbol{AS})) \quad (5)$$

where $\boldsymbol{W}$ is a positive T-by-F weight matrix and .* is the element-wise multiplication. The estimation of perceptually motivated weights is explained in this section, and the weighted NMF algorithm in next section.

A. Perceptually motivated weights

Function d (4) is linear as a function of the scale of the input, since $d(\alpha p, \alpha q) = \alpha d(p, q)$ for any positive scalar α. The divergence (3) is a sum of function d over $x_{t,f}$. Thus, the amount of contibution, or, quantitative "significance" of an individual observation $x_{t,f}$ in the divergence is $x_{t,f}$. Thus, the quantitative significance of an auditory object within a mixture in the parameter estimation is the sum of power spectral bins, which is the energy of the object.

The large dynamic range of the human auditory system is mainly caused by the non-linear response of the auditory cells, which can be modeled as a compression of the input signal separately within each critical band. In this system the compression is modeled by calculating a weight for each frequency bin in each frame. The weights are selected so that the weighted sum of spectrum bins is equal to the estimated loudness. This way the quantitative significance of a time-frequency component corresponds roughly to its "perceptual significance."

The loudness of one frame is modeled by calculating the excitation using perceptually motivated frequency scale 1/Bark and critical bandwidth [28, pp. 141-144], compressing the excitation, and integrating over frequency [28, p. 197]. Thus, the loudness can be estimated individually for each critical band. In our system, 24 separate bands are spaced uniformly on

the Bark scale and denoted by disjoint sets of frequency bins F_b, b = 1...24. The energy within each band in every frame is calculated as

$$e_{b,t} = \sum_{f \in F_b} x_{f,t} \tag{6}$$

The fixed power response of the outer and middle ear is taken into account by multiplying the energies within each band by the corresponding power response h_b. The energies after middle and outer ear filtering are given as

$$g_{b,t} = h_b e_{b,t} \tag{7}$$

To get a good match of loudness at the active regions of the input signal, h_b is set to be the equal-loudness contour at 60 dB [29, p. 55]. Term *loudness index* is used to refer to the the loudness estimate of a frame within a critical band. Loudness index $L_{b,t}$ of critical band b at frame t is given by:

$$L_{b,t} = [g_{b,t} + \varepsilon_b]^{\nu} - \varepsilon_b^{\nu} \tag{8}$$

where ν is a fixed compression factor, and ε_b is the threshold of hearing at band b. The loudness model of the system is adopted from the loudness models of Moore et al. [30] and Zwicker and Fastl [28, p. 201].

The threshold of hearing may not be known in practise, so it can be estimated from the input signal. The separation algorithm is noncausal so this is not a problem. For simplicity, ε_b is defined to be equal for all critical bands (the linear response of outer and middle ear is taken into account in h_b). The level of the signal is estimated from the variance σ^2 after middle and outer ear filtering, which is given as

$$\sigma^2 = \frac{1}{K} \sum_{b=1}^{24} g_{b,t} \tag{9}$$

A good choise for ε_b was found to be $10^{-5} \cdot \sigma^2$. For each critical band in each frame, weight $c_{b,t}$ is assigned, which mimics the loudness perception. The weights are selected so that the quantitative criterion, the weighted energy, equals the estimated loudness:

$$c_{b,t} e_{b,t} = L_{b,t} \tag{10}$$

from which $c_{b,t}$ can be solved as

$$c_{b,t} = \frac{L_{b,t}}{e_{b,t}} \tag{11}$$

From (7), (8) and (11) it can be seen that as the energy approaches zero, the limit for $c_{b,t}$ is:

$$\lim_{e_{b,t} \to 0} \frac{L_{b,t}}{e_{b,t}} = \nu h_b \varepsilon_b^{\nu - 1} \tag{12}$$

Therefore, for critical bands the energy of which is exactly zero, $c_{b,t}$ is set to $\nu h_b \varepsilon_b^{\nu - 1}$. For real-world signals the energy within a band will in practise never be exactly zero, but for generated test signals it is possible to have such situations.

Simulation procedures similar to those described in Section III were used to test different values of ε_b and ν. It was noticed that the algorithm is not sensitive for the exact values of the parameters. A good performance was obtained with $0 < \varepsilon_b \leq 10^{-5} \cdot \sigma^2$ and $0.17 < \nu < 0.3$. Too large ε_b or ν will decrease the amount of compression, which is the main motivation for using the weights.

To simplify the notation, let us denote the estimated weights by F-by-K matrix $[\boldsymbol{W}]_{ft} = c_{b,t}$, for all $f \in F_b$.

B. Algorithm for weighted non-negative matrix factorization

The estimation of the parameters $\boldsymbol{A}$ and $\boldsymbol{S}$ in (2) is done by minimizing the cost function (5) with respect to $\boldsymbol{A}$ and $\boldsymbol{S}$. The optimization algorithm updates randomly initialized $\boldsymbol{A}$ and $\boldsymbol{S}$ iteratively using multiplicative update rules. For the minimization of the unweighted divergence (3), Lee and Seung [11] proposed the multiplicative update rules

$$\boldsymbol{S} \leftarrow \boldsymbol{S}.*[\boldsymbol{A}^T(\boldsymbol{X}./\boldsymbol{A}\boldsymbol{S})]./[\boldsymbol{A}^T\boldsymbol{1}] \tag{13}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A}.*[(\boldsymbol{X}./\boldsymbol{A}\boldsymbol{S})\boldsymbol{S}^T]./[\boldsymbol{1}\boldsymbol{S}^T] \tag{14}$$

where .* and ./ are the element-wise multiplication and division, respectively, and $\boldsymbol{1}$ is an all-one N-by-F matrix. The divergence (3) was shown to be nonincreasing under the update rules [11].

For the weighted divergence (5) update rules for $\boldsymbol{A}$ and $\boldsymbol{S}$ are given as

$$\boldsymbol{S} \leftarrow \boldsymbol{S}.*[\boldsymbol{A}^T(\boldsymbol{W}.*\boldsymbol{X}./\boldsymbol{A}\boldsymbol{S})]./[\boldsymbol{A}^T\boldsymbol{W}] \tag{15}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A}.*[(\boldsymbol{W}.*\boldsymbol{X}./\boldsymbol{A}\boldsymbol{S})\boldsymbol{S}^T]./[\boldsymbol{W}\boldsymbol{S}^T] \tag{16}$$

which can be verified as follows. Let us write the weighted divergence in the form

$$D(\boldsymbol{X}||\boldsymbol{A}\boldsymbol{S};\boldsymbol{W}) = \sum_{f=1}^{F} D(\boldsymbol{x}^f||\boldsymbol{A}\boldsymbol{s}^f;\boldsymbol{w}^f) \tag{17}$$

where $\boldsymbol{x}^f$, $\boldsymbol{s}^f$, and $\boldsymbol{w}^f$ are the f^{th} columns of matrices $\boldsymbol{X}$, $\boldsymbol{S}$, and $\boldsymbol{W}$. The divergence for each frequency bin can be written as

$$D(\boldsymbol{x}^f||\boldsymbol{A}\boldsymbol{s}^f;\boldsymbol{w}^f) = D(\boldsymbol{D}^f\boldsymbol{x}^f||\boldsymbol{D}^f\boldsymbol{A}\boldsymbol{s}^f) \tag{18}$$

where $\boldsymbol{D}^f$ is a diagonal matrix in which the elements of $\boldsymbol{w}^f$ are on the diagonal. By assigning $\boldsymbol{D}^f\boldsymbol{x}^f = \boldsymbol{y}^f$ and $\boldsymbol{D}^f\boldsymbol{A}^f = \boldsymbol{B}^f$ the divergence (18) can be written as

$$D(\boldsymbol{D}^f\boldsymbol{x}^f||\boldsymbol{D}^f\boldsymbol{A}\boldsymbol{s}^f) = D(\boldsymbol{y}^f||\boldsymbol{B}^f\boldsymbol{s}^f) \tag{19}$$

for which the update rule (13) is given as

$$\boldsymbol{s}^f \leftarrow \boldsymbol{s}^f.*[(\boldsymbol{B}^f)^T(\boldsymbol{y}^f./\boldsymbol{B}^f\boldsymbol{s})]./[(\boldsymbol{B}^f)^T1] \tag{20}$$

where 1 is a all-one N-by-1 vector. By substituting $\boldsymbol{y}^f = \boldsymbol{D}^f\boldsymbol{x}^f$ and $\boldsymbol{B}^f = \boldsymbol{D}^f\boldsymbol{A}^f$ back to the equation, the update rule (20) can be written as

$$\boldsymbol{s}^f \leftarrow \boldsymbol{s}^f.*[\boldsymbol{A}^T\boldsymbol{D}^f(\boldsymbol{D}^f\boldsymbol{x}^f./\boldsymbol{D}^f\boldsymbol{A}\boldsymbol{s}^f)]./[\boldsymbol{A}^T\boldsymbol{D}^f1] \tag{21}$$

The above equals (15) for each column of $\boldsymbol{S}$, and therefore (15) is the update rule for weighted non-negative matrix factorization. Similarly, the update rule (16) for $\boldsymbol{A}$ can be verified by writing the weighted divergence individually for each row of $\boldsymbol{A}$.

As preprocessing, the power spectrogram $\boldsymbol{X}$ and the weight matrix $\boldsymbol{W}$ are calculated from the time-domain input signal using the procedure described in Section II. The number of components N is set by hand. N should be equal or larger than the number of clearly distinguishable musical instruments. For

a drum sequence, for example, one might use $N = 3$ for a drum pattern which consists of bass drum, snare, and hi-hat sounds.

The iterative algorithm is given as follows:

Step 1. Initialize each element of $\boldsymbol{A}$ and $\boldsymbol{S}$ with the absolute value of Gaussian noise.

Step 2. Update $\boldsymbol{S}$ using the multiplicative step (15).

Step 3. Update $\boldsymbol{A}$ using the multiplicative step (16).

Step 4. Evaluat the cost function and repeat steps 2 - 4 if needed.

The steps 2 - 4 are repeated until the value of the cost function does not change. In practise this is done by keeping track of the iteration steps for which the decrease of the cost function is smaller than a small threshold. Iteration is stopped when the decrease has been smaller for a certain number of iterations.

The computation time depends on the length and complexity of the input signal and on the number of components. For example, the separation of a 10-second polyphonic signal into four components takes about one hundred iterations to converge, which takes about half a minute on a regular desktop computer[1] when implemented in Matlab.

C. Synthesis

In the synthesis, the power spectrum of each component within every frame is calculated as $a_{t,n}\boldsymbol{s}_n$, $t = 1..K$, $n = 1..N$. The square root is taken element-wise to get the magnitude spectra. To get complex spectra, there are two alternative methods for the estimation of the phases. Either the phases of the original spectrogram can be used for the separated components, or the phase generation method proposed by Griffin and Lim [31] with the improvements proposed by Slaney, Naar, and Lyon [32] can be used.

In most cases where the separation is successful the use of original phases produces good results. It also allows the synthesis of sharp attacks with an accuracy which would otherwise be impossible with 40 ms window sizes. However, if the original phases are for some reason not suitable for the separated magnitude spectrogram, the resulting time-domain signal may become distorted because of discontinuities at at the frame boundaries.

In our simulations, the best perceptual quality was obtained using the original phases and the following overlap-add procedure. First, the complex spectrogram of a component is obtained by assigning the original phases for the separated spectrogram. Second, the time-domain signal of each frame is obtained by the inverse discrete Fourier transform. Third, each frame is windowed using the square root of the Hanning window. Finally, the frames are concatenated using overlap-add. The windowing eliminates most discontinities between frames. Since the square root of Hanning window is used twice, both in the analysis and in the synthesis, the method allows perfect reconstruction since adjacent windows sum to unity. This method was found to produce the best quality, and also the computational cost is very low. The described synthesis method was used for all algorithms in the simulation experiments that are presented in the next section.

1 1.7 GHz Pentium 4

Table I: Description of the test categories. For each category, 150 signals with randomly selected samples were generated.

category	description
1	Two equal-length pitched samples which overlap half of their duration; the second one sets on at the half of the duration of the first one.
2	Two equal-length pitched instrument samples which set on simultaneously and overlap their whole duration.
3	Two percussive sounds. The first one is repeated alone three times, then both sounds are repeated three times simultaneously, and in the end the second sound is repeated three times. The repetitions are random instants of both classes, i.e. not identical samples. There is a 200 ms interval between the repetitions.
4	A pitched instument sample and a percussive sound which is repeated six times. The repetitions are random instants of a class, with a 200 ms interval between the repetitions. The first repetition onsets simultaneously with the pitched sample. The number of overlapping repetitions depends on the length of the pitched sample.
5	Two pitched instrument samples and two percussive sounds (the sum of signals from categories 1 and 3)

III. SIMULATION EXPERIMENTS

In general, the objective of sound source separation algorithms is to extract sound sources which are perceptually close to the original ones before mixing. Quantitative evaluation of the perceptual separation quality is difficult. It can be measured either by listening tests or by computational procedures which compare the ideal source signals with separated signals. Basically in both cases, the source signals before mixing are required. In practise this will limit us to synthesized test signals.

A. Signals

Since the performance of all one-channel data-driven separation algorithms is currently very limited and no systematic evaluation of the algorithms has been published, the evaluation was performed using simple “elementary” mixing situations of pitched and percussive sounds. Five different test categories were used and 150 test cases were generated within each category. The categories are described in Table I.

The samples of pitched instruments were randomly drawn from a database of 4379 samples of individual notes. The database is a combination of samples from the McGill University Master Samples Collection [33], the University of Iowa website [34], IRCAM Studio Online [35]. The total number of pitched instrument samples was 4378, each having the sampling frequency 44100 Hz..

The percussive instrument database is recorded at the Tam-

pere University of Technology and it consists of 48 classes of human-made percussive sounds such as clapping of hands and tapping of foot. For each class there are 15 instances, which are perceptually similar but the acoustic waveform may differ.

For each category, 150 mixtures were generated. In the categories 1, 2, 4, and 5, the pitched instument samples were randomly drawn from the database. The samples within a mixture were truncated so that the lengths become equal. However, in all test cases the maximum length of the samples was limited to two seconds. The fundamental frequencies or instruments within each mixture were not allowed to be equal. In percussive mixtures (categories 3, 4, and 5) a random class of sounds was selected for each percussive source. For each source, six instances were randomly drawn from the class, one for each repetition.

To simulate the mixing conditions encountered in real-world situations, the dynamic difference between sources is varied. For each pair of samples, the difference was randomily drawn from a uniform distribution between -10 and 10 dB, and the samples were scaled to obtain the desired power ratios. The samples were summed and the mixture signal is normalized to range [-1, 1]. The original sources before mixing were stored to allow the measurement of the separation quality.

B. Algorithms

In addition to the proposed method, some recently published data-driven algorithms were used as a baseline in the systematic evaluation. All the algorithms use the same preprocessing that was described in the beginning of Section II and synthesis based on the original phases as described in Section C. However, all the other algorithms except the proposed one performed better when magnitude spectrogram was used as an observation. Therefore, magnitude spectrograms were used for all the other algorithms except the proposed one. For all the algorithms the number of components was set to equal the number of sources in the mixtures. The following algorithms were tested:

1) Independent subspace analysis. The implementation of the algorithm follows the outline proposed by Casey and Westner [4], using which the algorithm was implemented. Two different algorithms for the estimation of independent spectra were tested: FastICA [23], [24] and Jade [25], [26]. However, differences between the algorithms were very small, so the results are shown only for Jade, which performed slightly better.

2) Non-negative matrix factorization was tested with the algorithms proposed in [11]. These minimize the unweighed divergence or the euclidean distance, and are denoted by NMF-DIV and NMF-EUC, respectively. The algorithms were implemented using the reference.

3) Non-negative sparse coding. The algorithm is a combination of multiplicative update and projected steepest descent, and it follows the outline proposed in [19]. Since the algorithm allows the usage of perceptually motivated weights as proposed in [21], the parameters are estimated using the weighted Euclidean distance. Also cost terms which favours the sparseness of the mixing matrix and the temporal continuity of mixing weights can be used. Unlike in the reference [19], the squared difference between the weights of adjacent frames was used as a cost function to obtain temporal continuity, since it produced better results. The algorithm in which sparseness is favoured is denoted by SC, and an algorithm in which temporal continuity is favoured is denoted by SC+temp.

4) The proposed method, which is denoted as W-NMF.

C. Evaluation

The evaluation can be done by comparing the separated components with the original sources which were stored. First, each separated component has to be assigned to a source. A separated component with the time-domain signal $\hat{u}(t)$ is clustered to source $u_m(t)$ which minimizes the residual-to-signal ratio (RSR)

$$RSR(m) = \frac{\sum_t [u_m(t) - \hat{u}(t)]^2}{\sum_t u_m(t)^2} \qquad m = 1 \ldots M \qquad (22)$$

where M is the number of sources. The use of the RSR proved to be a robust way of assigning separated components to sources. The algorithms may use more than one component to represent a source. In this case the components assigned to the source are summed. For algorithms which estimate the parameters using the magnitude spectrogram, the summation can be done for the synthesized time-domain signals. For the proposed method the summation has to be done for the estimated power spectrograms of the components, which are then resynthesized.

If no components are assignined to a source, the source is said to be undetected. This is defined to be a *detection error*. The *detection error rate* is defined to be the ratio of the total number of undetected sources to the total number of sources. It measures how well an algorithm is able to detect sources from a signal.

The perceptual audio quality measure (PAQM, [36]) is used to evaluate the perceptual quality of separated signals. The PAQM is a computational procedure which can be used to estimate the perceptual difference between two audio signals. In the PAQM, both signals are processed using a specific auditory model, and the difference in the auditory domain is calculated. The difference is called noise disturbance, and it is usually examined on the logarithmic scale. In our case the PAQM is calculated between the separated and the original signals. The measure is calculated for sources for which at least one component is assigned. The PAQM algorithm was implemented according to the reference [36].

In addition to PAQM, the quality is also evaluated by calculating the RSR (22) between the separated and the original signal. In this case the residual is the error between the separated signal and original one, and RSR is the ratio between the energies of the error signal and original signal. For both measures, the averaging over test cases is done before taking the logarithm. The measures are not calculated for undetected sources.

D. Results

The obtained results are illustrated in Fig. 1. Each measure in categories 1 to 5 is an average of 150 test cases. Column 'all' is the average of all the categories. The PAQM scale can be roughly interpreted so that measures below -1.5 are of a relatively high quality, and measures above -0.5 mean large distortions. Because of the logarithmic scale, even small changes in the PAQM mean large differences in the quality. RSR measures the average amount of interference in the separated signals. The differences between PAQM and RSR are large, which is natural since the way the measures are calculated are totally different. The PAQM measures better the quality at low-intensity regions, since it uses an auditory model which compresses the signals.

The differences between the categories are large. Category 1 was found to be a fairly easy separation task, since both samples occur also separately in the mixture. Category 2 seemed to be the most difficult one. The samples overlap their whole duration and if the amplitude envelopes are similar, the separation is very difficult, since the algorithms do not utilize the harmonic structure of the sources.

According to the detection error rate, the proposed method (W-NMF) produced good results in all categories except the second one, in which all the algorithms had big difficulties. The overall detection rate was best for the proposed method, even though the differences between the algorithms are quite small.

The PAQM indicates that the proposed method outperforms all the other algorithms clearly in the perceptual quality. Thanks to the perceptually motivated weights, W-NMF models the low-intensity sections more accurately than the other algorithms. According to the RSR, the differences between the algorithms are small, and none of the algorithms performs clearly better than the others. To obtain better accuracy at low-intensity sections, the proposed method tends to allow errors on high-intensity sections. Therefore, the performance is only moderate according to the RSR.

The performance of the sparse coding algorithms is comparable with that of the NMF. In general the use of the sparseness term does not improve the quality, but the use of the temporal continuity constraint improved the quality of pitched sounds, so that the best performance among non-ISA algorithms in category two was obtained with SC+temp.

Informal listening test showed that the perceptual quality of

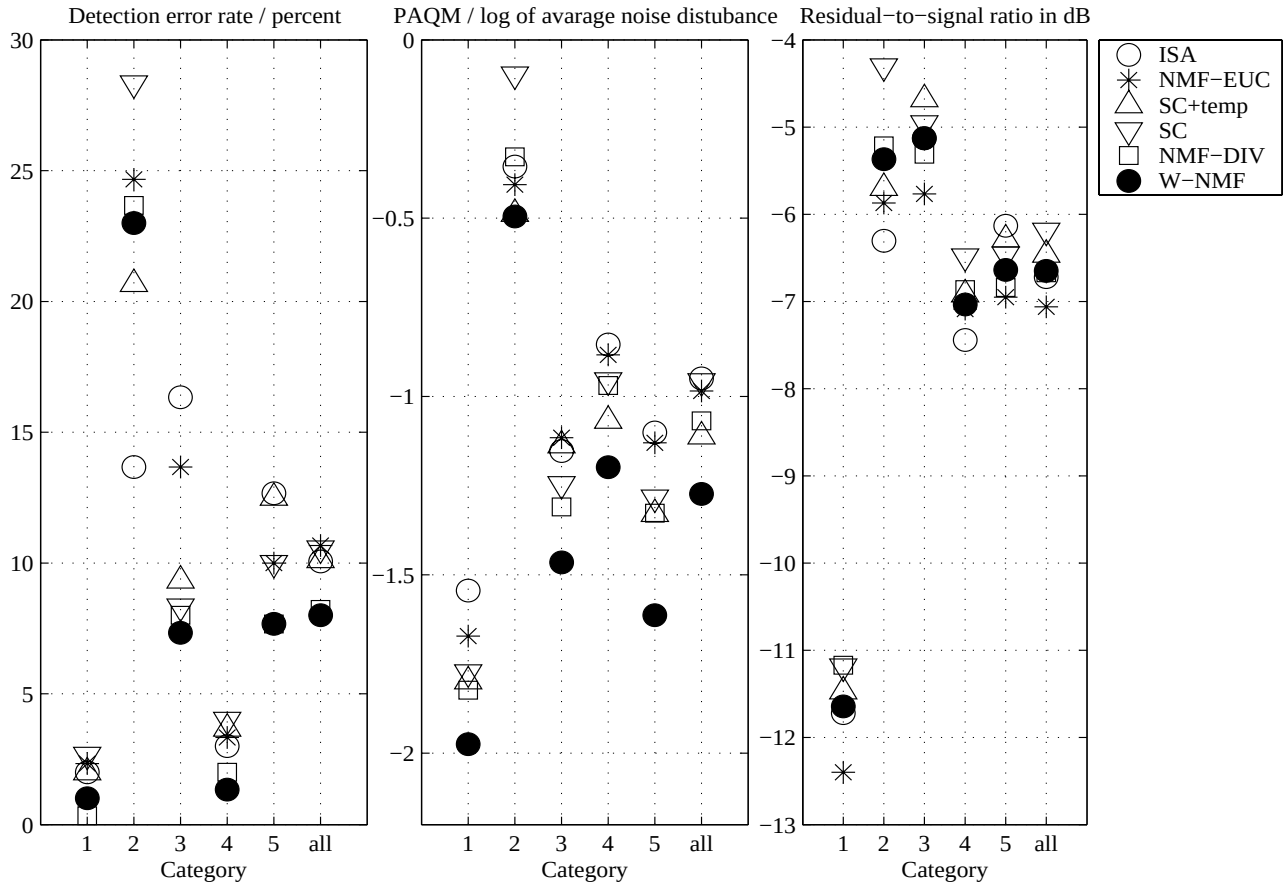

Fig. 1. The average detection error rates (left plot), perceptual audio quality measures (middle plot), and residual-to-signal ratios (right plot) in different test categories. The acronyms for different algorithms are explained in the text. For all measures, the smaller the measure, the better the quality. For categories 1 to 5, the measure is an average of 150 mixtures, which contain either two or four sources. Category 'all' is an average of categories 1 to 5.

synthesized sources correlated well with the PAQM. For a human listener all the cases can be considered to be relatively easy separation tasks. The humans' ability to utilize the harmonic structure of sounds makes the separation of harmonic sounds fairly easy task even in category 2 which was most difficult for tested algorithms. In informal listening test the categories 3 and 5 were found to be most difficult ones, since the repetition rate was identical for both percussive sources.

E. Convolutive models or multiple components per source

The limitations of the signal model can be overcome either by using multiple components per source or a more complex signal model. When multiple components per source are used, the components has to be clustered into sources. In general this is a difficult task. Some clustering principles have been proposed by e.g. Casey and Westner [4] and by Virtanen [19].

The presented simulation experiments were carried out also with multilple components per source and with convolutive models. The convolutive models and their estimation algorithms are derived from NMF and non-negative sparse coding. The algorithm proposed by Smaragdis [20] is denoted by NMD and the algorithm proposed by Virtanen [21] is denoted by CONV-SC. NMD was implemented using the reference, with the exception of new update rules, which update all the spectra simultaneously, which makes the algorithm more efficient.

Since the optimal number of components is not known, ISA was tested with 6, 16, and 40 components, referred as ISA(6), ISA(16), and ISA(40). The Jade algorithm was computationally too heavy with 40 components, so FastICA was used as ISA(40). The proposed method (W-NMF) was tested with 6 and 16 components, denoted by W-NMF(6) and W-NMF(16), respectively.

In our simulations the objective was to compare separation algorithms, not clustering algorithms, so the clustering was avoided by using the original sources as a reference. The components were assigned to sources using the RSR (22) as described earlier. This method was found to be a very robust way of clustering components to sources also with multiple components per source. Using this procedure a performance measure for the multi-component algorithms in the ideal case was obtained; in practise it will be very difficult to get perfect clustering.

The results are presented in Table II. For both ISA and W-NMF, the detection error rate decreases when the number of components increases. For ISA also the quality of separated signals increases with multiple components. However, for W-NMF the perceptual quality of separated signals does not increase with multiple components. The highest perceptual quality of separated samples among all the algorithms is obtained with W-NMF when the number of components equals the number of sources.

The performance of the convolutive models (NMF and CONV-SC) is worse than the proposed method, but comparable with the other algorithms. The algorithms are based on the assumption of repetitive sources, and for the test signals this holds only for the percussive sources. The algorithms may have potential in source separation, but that is out of the scope of this paper.

Table II: Simulation results for algorithms with multiple components per source and convolutive models. The measures are the average of test categories 1 to 5.

algorithm	Detection error rate	PAQM	RSR
ISA	10.1	-0.95	-6.7
ISA(6)	1.7	-1.07	-7.5
ISA(16)	0.3	-1.13	-7.7
ISA(40)	0.1	-1.18	-7.8
W-NMF	8.0	-1.27	-6.7
W-NMF(6)	1.8	-1.25	-7.1
W-NMF(16)	0.6	-1.23	-7.4
CONV-SC	11.3	-1.08	-6.5
NMD	11.3	-1.02	-6.4

F. Discussion

The only category in which ISA performs systematically better than algorithms based on the NMF is the second one, in which the samples overlap their whole duration. This suggests that the independence of spectra is a better separation criterion if the objective is to separate sources which are overlapping their whole duration. If two sources are always present simultaneously, it might even be desirable to model them as one source. The performance of all the algorithms was poor in the category two. In human sound perception the harmonicity is a very strong grouping principle. However, it will be very difficult to incorporate harmonicity restriction to any of the tested algorithms, since they use the DFT spectrum to characterize a source. Harmonicity will require the paramterization of the fundamental frequency, and in practise a parametric source model has to be used.

In the proposed method the number of components has to be set by hand. Currently there is no method for the automatic estimation of the number of components. In practise this can be solved by using a large number of components, which are then clustered to sound sources. If the number of components is large, the estimation algorithm becomes computationally very slow. This makes the estimation algorithm unpractical for long signals. This can be solved by analysing the input signal in short segments.

The proposed separation algorithm was also tested using polyphonic music signals. Demonstration signals are available at http://www.cs.tut.fi/~tuomasv/.

IV. CONCLUSIONS

A data-adaptive algorithm for the separation of sound

sources from one-channel signals is presented. The existing algorithms based on the linear generative model are limited in a sense that the accurate summation of sources in the power spectral domain and a large dynamic range may not be obtained simultaneously. In practise, high-intensity observations will dominate the separation, and a good perceptual quality is difficult to obtain since also low-intensity observations are perceptually significant. The use of the weighted cost function is a simple and efficient way for obtaining a large dynamic range in this framework. Perceptually motivated weights can be used to approximate the loudness perception of the human auditory system. The parameters of the sources can be efficiently estimated using the proposed weighted non-negative matrix factorization algorithm.

Simulation experiments indicate that the proposed method enables higher perceptual quality of separated components than the existing algorithms, e.g. widely used independent subspace analysis. In some cases it is possible to obtain a relatively high separation quality. However, none of the tested algorithms was able to separate robustly two-note mixtures of pitched instrument samples, which overlap their whole duration. This suggests that data-driven algorithms may not be sufficient for the separation of music signals without a prior knowledge, such as harmonicity.

ACKNOWLEDGEMENT

The database of percussive samples was recorded by Jouni Paulus. The author would like thank Anssi Klapuri and Jouni Paulus for helpful discussions and constructive criticism when preparing this paper.

REFERENCES

[1] M. Goto, “A Predominant-F0 Estimation Method for Real-world Musical Audio Signals: MAP Estimation for Incorporating Prior Knowledge about F0s and Tone Models,” in *Proc. of Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis*, 2001.

[2] D. L. Wang, and G. J.Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. on Neural Networks*, vol 10. pp. 684-697, 2001.

[3] A. L. C. Wang, “Instantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation,” Ph.D. dissertation, Stanford University, 1994.

[4] M. A. Casey, and A. Westner, “Separation of mixed audio sources by independent subspace analysis,” in *Proc. International Computer Music Conference*, 2000.

[5] D. FitzGerald, E. Coyle, B. Lawlor, “Sub-band independent subspace analysis for drum transcription,” in *Proc. of the 5th Int. Conference on Digital Audio Effects*, 2002.

[6] C. Uhle, C. Dittmar, T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

[7] I. Orife, “Riddim: A Rhythm Analysis And Decomposition Tool Based On Independent Subspace Analysis,” M.A. thesis, Dartmouth College, 2001

[8] E. Vincent and X. Rodet. “Music transcription with ISA and HMM,” In *Proc. of the 5th Int. Symposium on ICA and BSS*, 2004.

[9] J. C. Brown, and P. Smaragdis, “Independent component analysis for automatic note extraction from musical trills,” *J. Acoust. Soc. Amer.* vol. 115, no. 5, May 2004.

[10] M. Casey, “MPEG-7 sound-recognition tools,” *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 11, no. 6, June 2001

[11] D. D. Lee, and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Advances in Neural Information Processing*, vol. 13, MIT Press, 2001.

[12] P. Smaragdis, and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[13] Y.-C. Cho, S. Choi, S.-Y. Bang, “Non-negative component parts of sound for classification,” in *Proc. IEEE Int. Symp. Signal Processing and Information Tech.*, 2003.

[14] B. A. Olshausen, and D. F. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” *Vision Research*, vol. 37, no. 23, pp. 3311-3325, 1997.

[15] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, “Automatic music transcription and audio source separation,” *Cybernetics and Systems*, vol. 33, no. 6, pp. 603-627, 2002.

[16] S. A. Abdallah, “Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models,” Ph.D. disseration, Department of Electronic Engineering, King’s College London, 2002

[17] L. Benaroya, F. Bimbot, L. McDonagh, and R. Gribonval. “Non negative sparse representation for Wiener based source separation with a single sensor,” in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, 2003.

[18] P. Hoyer. “Non-negative sparse coding,” *Neural Networks for Signal Processing XII*, 2002.

[19] T. Virtanen, “Sound source separation using sparse coding with temporal continuity objective,” in *Proc. International Computer Music Conference*, 2003.

[20] P. Smaragdis, “Discovering auditory objects through non-negativity constraints,” in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004

[21] T. Virtanen, “Separation of sound sources by convolutive sparse coding,” in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004

[22] M. D. Plumbley, and E. Oja, “A ’non-negative PCA’ algorithm for independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 15, no. 1, pp 66-76, 2004.

[23] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. on Neural Networks* vol. 10, no. 3, pp. 626-634, 1999.

[24] “FastICA package for MATLAB,” http://www.cis.hut.fi/projects/ica/fastica/

[25] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Computation* vol 11, no. 1, pp. 157-192, 1999.

[26] J.-F. Cardoso, “Jade algorithm for Matlab”, http://www.tsi.enst.fr/icacentral/algos.html, March 17, 2004.

[27] D. FitzGerald, “Automatic Drum Transcription and Source Separation,” Ph.D. dissertation, Dublin Institute of Technology, 2004

[28] E. Zwicker, and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, Berlin, 1999. Second Edition.

[29] B. C. J. Moore, An Introduction to the Psychology of Hearing, , Academic Press, 2000. 4th edition.

[30] B. C. J. Moore, B. Glasberg, and T. Baer, “A model for the prediction of thresholds, loudness, and partial loudness,” *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224--240, 1997.

[31] D. Griffin, and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32, pp. 236-242, 1984.

[32] M. Slaney, D. Naar and R.F. Lyon. “Auditory model inversion for sound separation,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1994, vol. II, pp.77-80.

[33] F. Opolko, and J. Wapnick, McGill University Master Samples (compact disk). McGill University, 1987.

[34] The University of Iowa Musical Instrument Samples Database, http://theremin.music.uiowa.edu.

[35] IRCAM Studio Online, http://soleil.ircam.fr/

[36] J. Beerends, and J. Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sounds presentation,” *J. Audio Eng. Soc.*, vol. 40, no. 12, 1992.