Iterative Monaural Audio Source Separation for Subspace Grouping

Martin Spiertz and Volker Gnann
Institut für Nachrichtentechnik
RWTH Aachen University
Tel: ++49 241 8027676, Fax: ++49 241 8022196
E-mail: {spiertz, gnann}@ient.rwth-aachen.de

***Abstract*—Monaural blind audio source separation usually separates a mixture into more signals than active sources. Therefore, a clustering of the separated signals is needed to reconstruct the sources. We propose a new iterative clustering and show that this approach outperforms classical clustering approaches which use features of the separated signals for clustering. The iterative clustering starts with the separation into two source estimates. Based on this, at each iteration the squared error between the source estimates of the former iteration and a linear superposition of the separated signals of the current iteration is minimized. The corresponding linear superposition generates new source estimates. The algorithm is evaluated on a large test set regarding melodies of different instruments, singing, and speech from the EBU.**

I. Introduction

A human listener usually observes several active sound sources at the same time. The listener is able to concentrate on one active source by suppressing the others. Blind audio source separation tries to emulate this human ability in an automatic way. Generally, audio source separation is useful in each case when a signal is distorted by other signals or simply by noise; e.g. in the case of unwanted instruments in a recording [2].

A general approach for blind audio source separation can be divided into four steps [7]. First the mixture is splitted into temporal segments, because stationarity is assumed for the separation process. Then, the temporal segments are separated into several signals called channels. The separation is done by the independent subspace analysis (ISA). If there are more channels than active sources, the channels are clustered to create as many source estimates as active sources. Finally, the source estimates need to be classified to reconstruct the separated sources over several temporal segments. In this paper, we address the problem of clustering the separated channels in the case of a single mixture. In [3], [7], and [10], several clustering methods based on signal features were proposed for the ISA. We will show, that a stable clustering algorithm for a large range of test signals is still an open problem, and introduce a clustering approach that outperforms the other tested clustering strategies.

This paper is structured as followed. In Section II the basic terms and algorithms are introduced. In III, the proposed clustering algorithm is outlined. The experimental setup and the corresponding results are presented in IV, followed by the conclusions in Section V.

II. Fundamentals

A. Signal Model

In this paper, we consider only sampled signals. We assume two active sources $s_k(t)$, $k\epsilon\{1,2\}$ observed with one sensor. We assume instantaneous mixing without any reverberation. For this model, the mixture $x(t)$ is represented by the addition of the two input signals,

$$x(t) = s_1(t) + s_2(t)\ . \tag{1}$$

The mixture $x(t)$ is transformed into time-frequency domain by the short-time Fourier transform (STFT). A separation algorithm separates the magnitude spectrogram $|\mathbf{X}|$ into N real-valued spectrograms $\mathbf{C}_n$, $1 \leq n \leq N$, see Section II-B. The time signals $c_n(t)$ are called channels. They are evaluated by multiplication of the corresponding spectrograms $\mathbf{C}_n$ with the phase of $\mathbf{X}$ and applying the inverse STFT. Usually the mixture is over-separated ($N > 2$), so the channels need to be clustered to two source estimates $y_k(t)$. Finally, the source estimates have to be assigned to the two active sources by solving the permutation of the two cluster outputs towards the two sources.

B. Separation Algorithm

ISA for monaural source separation was first introduced in [2]. In [8], the algorithm is explained in detail. The ISA decomposes the magnitude spectrogram into a set of frequency basis vectors $\mathbf{f}_n$, and a set of corresponding time varying gains $\mathbf{g}_n$ using a singular value decomposition (SVD). For the ISA, it is assumed that the components of the spectrogram of two sources are statistical independent. Maximizing the statistical independence between the gain vectors is proposed in [2]. As mentioned in [9], maximizing the statistical independence between the frequency basis vectors leads to better results. We will use this version of the ISA for our experiments. A standard algorithm for maximizing the statistical independence for a set of vectors is the independent component analysis [1]. The real valued spectrograms of the channels are evaluated by multiplying the basis vectors with their gains,

$$\mathbf{C}_n = \mathbf{f}_n \mathbf{g}_n^T\ . \tag{2}$$

By dropping small eigenvalues of the SVD, a dimension reduction is introduced, which leads to the number N of separated channels c_n evaluated by the ISA. Because of this dimension reduction, the sum of all channels generally is not equal to the mixture,

$$\sum_{n=1}^{N} c_n(t) \neq x(t) \; . \tag{3}$$

For our experiments we require the sum of all channels to be equal to the mixture, as further explained in Section III. In this case the separation can be done for $N-1$ channels and the last channel c_N is defined as the error signal,

$$c_N(t) = x(t) - \sum_{n=1}^{N-1} c_n(t) \; . \tag{4}$$

C. Clustering Algorithms

In [10], several strategies are introduced for clustering the independent channels of the ISA. The main concept for these clustering algorithms is the fact that some channels share certain features with the source signal (e.g. the envelope in time or frequency domain). Therefore a clustering by that features is possible. For our experiments, we have implemented the *agglomerative clustering* and the *regrouping algorithm*. In both cases, we use the Kullback-Leibler distance (KLD) as the distance measure [10]. Three features are used for clustering. These are a) the modified envelope proposed by [10], b) the absolute value of the Fourier transform, and c) the audio spectrum envelope (ASE) as proposed in the MPEG-7 framework [6].

The distances of the different channels has to be evaluated for clustering. We call a matrix with the distance between the channels $c_i(t)$ and $c_j(t)$ at row i and column j distance matrix. In [3] a *clustering by friends* is introduced which can be applied to arbitrary distance matrices. The algorithm treats each column of the distance matrix as a probability distribution. At each iteration, the entry of the new distance matrix at row i and column j is evaluated as the distance between the columns i and j. Thereby, channels sharing large (respectively small) distances to the same group of other channels are rated as very similar. With this iteration scheme the algorithm converges towards a binary matrix. Channels with identical columns in the final distance matrix are assumed to belong to the same source. The channels of the ISA are assumed to be uncorrelated. Therefore the higher order cumulants of the channels are used as statistically motivated features for initializing the distance matrix [3]. The distance measure is again the KLD. Additionally, we have also evaluated distance matrices based on distances between envelopes as defined in [10].

Finally in [7] a *k-means clustering* with the KLD is introduced for clustering the frequency basis vectors $\mathbf{f}_n$.

D. Performance Measurement

Performance measurement for audio source separation can be done by the measures Source to Distortion Ratio (SDR_1), Source to Interference Ratio (SIR), and Source to Artifacts Ratio (SAR) [5]. For these distortion measures, the error between the input and the output of an arbitrary source separation algorithm is separated into an error by interference of other sources (leading to the measure SIR) and the remaining error based on artifacts (leading to the measure SAR). Because many source separation algorithms reconstruct the original signals up to a permutation and a multiplicative gain, these measures evaluate an optimal gain for minimizing the appropriate distortions. The ISA separates the sources without any multiplicative gains. Therefore distance measures like the improvement of the signal-to-noise ratio (ISNR) for source k [7],

$$\text{ISNR}_k = 10\log_{10}\frac{\sum_t (x(t)-s_k(t))^2}{\sum_t (y_k(t)-s_k(t))^2}[\text{dB}] \; , \tag{5}$$

or the SDR_2 based on the magnitude spectrogram can also be used for the comparison of the results [9]. The indices for $\text{SDR}_{1,2}$ are introduced to distinguish the two variants.

III. Proposed Clustering Algorithm

In the following we will drop the time index t, if there is no possibility for confusion. For finding the optimal clustering strategy in the meaning of ISNR we try to maximize the mean ISNR over both sources, which can be expressed as,

$$\frac{1}{2}\sum_{k=1}^{2}\text{ISNR}_k = 5\log_{10}\frac{\sum_t (x-s_1)^2 \sum_t (x-s_2)^2}{\sum_t (y_1-s_1)^2 \sum_t (y_2-s_2)^2}[\text{dB}] \; . \tag{6}$$

Because the numerator in (6) is constant for arbitrary clustering, maximizing the mean ISNR leads to minimizing the cost function e,

$$e = \sum_t (y_1-s_1)^2 \sum_t (y_2-s_2)^2 \; . \tag{7}$$

In general, the y_k can be described as

$$y_1 = \sum_{n=1}^{N} a_n c_n \; , \tag{8}$$

$$y_2 = \sum_{n=1}^{N} (1-a_n) c_n \; , \tag{9}$$

and clustering can be defined as the process of finding optimal weights a_n. Minimizing the cost function in (7) over a_n leads to the derivative,

$$\begin{aligned}\frac{de}{da_n} = &\sum_t 2c_n(y_1-s_1)\sum_t (y_2-s_2)^2 \\ &- \sum_t 2c_n(y_2-s_2)\sum_t (y_1-s_1)^2 \; .\end{aligned} \tag{10}$$

With (4) we set (10) to zero and get the following result:

$$\sum_t c_n s_1 = \sum_{i=1}^{N} a_i \sum_t c_n c_i \; . \tag{11}$$

The derivation of all a_n lead to the linear system of equations defined by

$$\mathbf{Aa} = \mathbf{b}\,, \tag{12}$$

$$\mathbf{A}[n,m] = \sum_t c_n(t)c_m(t)\,, \tag{13}$$

$$\mathbf{b}[n] = \sum_t c_n(t)s_1(t)\,, \tag{14}$$

$$\mathbf{a}[n] = a_n\,. \tag{15}$$

Because the s_k are unknown for blind source separation tasks, we start with separating x into two channels. In this case, no clustering needs to be done, and the two channels correspond to the source estimates y_k. Now the iterative clustering can be described by separating the mixture into N channels and replacing the input signal s_1 in (14) by the source estimate y_1,

$$\mathbf{b}[n] = \sum_t c_n(t)y_1(t)\,. \tag{16}$$

Evaluating the optimal weights a_n in (12) leads to new source estimates y_k as defined in (8) and (9).
If this iterative process is done for all N up to a maximum number of channels $N_{\max}$, we call it *N-step clustering*. By evaluating an ISA at each iteration, the N-step clustering introduces a high computational complexity to the separation process. Therefore we also evaluate a *two-step clustering* algorithm. For this algorithm, we perform only the first separation into two channels followed by a second separation directly into $N_{\max}$ channels. With (12) the $a_n \epsilon \mathbb{R}$ are not restricted, which we will call the *unconstrained* approach. By interpretation of the a_n as weights for a clustering with the maximum value of 1 (channel c_n belongs complete to source estimate y_1) and the minimum value of 0 in the opposite case, we get the constraints $0 \leq a_n \leq 1$. If the a_n are restricted by these constraints the clustering will be called *constrained*.

IV. Experimental Results

The test set consists of each of the 23 melodious phrases plus the soprano and the tenor singer and the English female/male speech of the Sound Quality Assessment Material from the EBU [4]. These are 27 input signals of roughly 10 seconds length each. All possible combinations lead to 351 mixtures. For stereo signals only the first channel of the stereo signal is kept. The ISA assumes stationary signals. Therefore we split these mixtures temporally into segments of 0.25 s. The whole source separation process is applied to each of these segments, as described in Section II. For the STFT, a window length of 1024 samples with a Hann window and 50% overlap is applied. After separation and clustering, the two source estimates y_k are concatenated with knowledge of the source signals s_k for generating the output signals. This is done because we consider only the clustering of the channels in this paper. The mean values over both sources are used for performance evaluation. For the *k-means clustering* the parameter η is set to 0.1 and the two cluster means are initialized as the two vectors with highest energy.

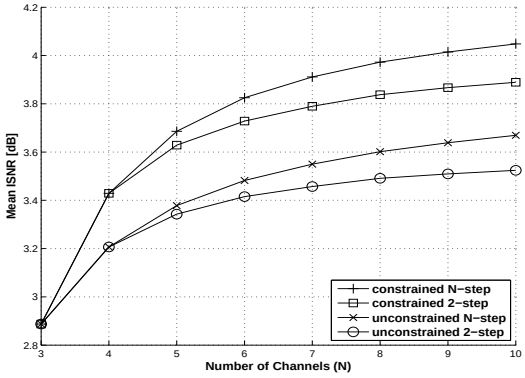

Fig. 1. Mean ISNR over all mixtures for the iterative clustering strategies plotted over the number of channels.

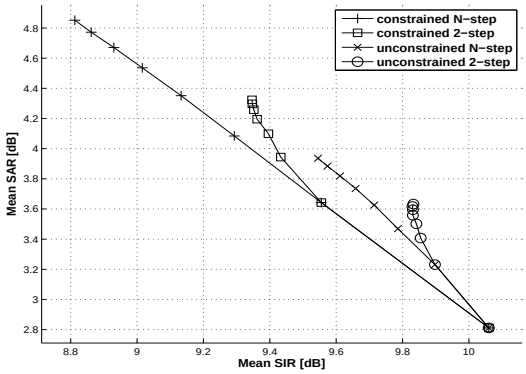

Fig. 2. Mean ISNR plotted over mean SIR over all mixtures for the four iterative clustering strategies. The number of channels N increases from the bottom right to the upper left marker.

In Figure 1, the mean ISNR over all 351 mixtures is plotted for the four different iterative clustering strategies outlined in Section III. Since the results for the SDR_1, SDR_2, and the SAR are similar, we will concentrate in the following on the ISNR. As can be seen from Figure 1 the constrained approach leads to better results than the unconstrained approach in terms of SDR_1, SDR_2, SAR and ISNR. Hence we will consider only the constrained approach in the following. The N-step clustering gives the best results, however the two-step clustering is approximately 2.8 times as fast as the N-step clustering. We will consider both in the following, showing the trade-off between computational complexity and ISNR.
The SIR behaves differently compared to SDR_1, SDR_2, SAR and ISNR, which is shown in Figure 2. Here, the SAR is plotted over the SIR. By increasing the number of channels the SIR decreases and the SAR increases. This trade-off between SAR and SIR exists for the four iterative grouping strategies. In general, the unconstrained approach leads to higher SIR values than the constrained approach. Also the two-step clustering leads to higher SIR than the N-step clustering. It can be seen that each approach with higher SIR values leads

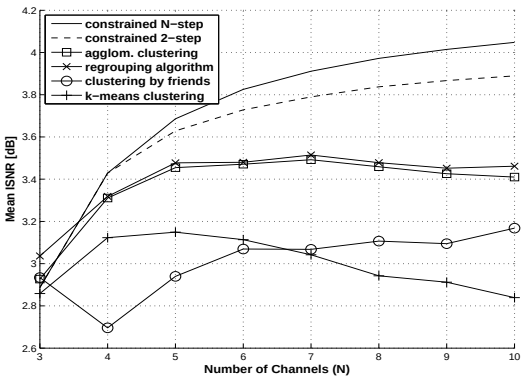

Fig. 3. Mean ISNR for all mixtures plotted over the number of channels for the evaluated clustering strategies.

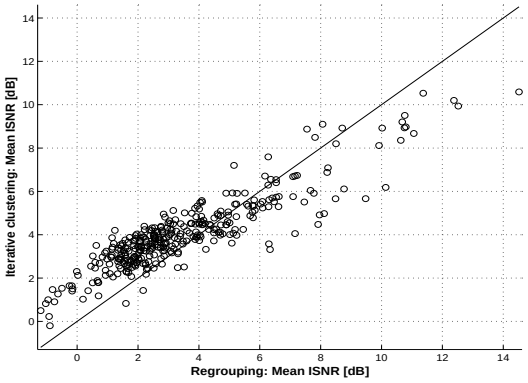

Fig. 4. ISNR for the *regrouping algorithm* based on audio spectrum envelope vs. iterative clustering. Each marker in the plot corresponds to one mixture. The diagonal line corresponds to the points of equal separation quality.

to lower SAR values. The SAR values are typically several dB below the SIR values. This is due to the fact that the artifacts are more dominant than the interferences. Therefore, even with decreasing SIR the increasing SAR also leads to increasing ISNR, SDR_1, and SDR_2, and in consequence, to an increasing reconstruction quality.

In Figure 3, the mean ISNR over all mixtures for the two constrained iterative clustering methods is compared with the clustering approaches of [3], [7], and [10]. From the different clustering approaches based on [10], the ASE as clustering feature leads to the best results and is selected for comparison. The *clustering by friends* is performed on a distance matrix filled with the distances between the spectral envelopes because this leads to better results than the statistically motivated distances between the cumulants of the channels. In [3] it is mentioned that *clustering by friends* needs more than 40 channels for working effectively. This explains the poor performance for a few channels as shown here. For all clustering strategies, only the iterative clustering improves the ISNR for increasing number of channels N. The main problem for feature-based clustering approaches proposed by [3], [7], and [10] is the increasing degree of freedom for increasing N. This is due to the fact that the number of possible clusterings of N channels to two source estimates is 2^N. Also, the higher number of channels leads to many channels with weak features, and therefore result in high failure probabilities.

As we can see, the clustering algorithms proposed by [10] perform very well for the chosen test set and testing conditions. In Figure 4 the mean ISNR of the *regrouping algorithm* vs. the constrained N-step clustering is compared for $N = 10$. It can be seen, that the distribution of the markers roughly follows the line of equal separation quality. For mixtures that are easy to separate (plotted in the upper right corner) the *regrouping algorithm* leads to higher ISNR values. This is due to the fact that in the case of good separation quality, the channels share most of the features of the underlying source, such that a feature-based clustering can be applied. For the major parts of mixtures with low separation quality (plotted in the lower left corner) the iterative clustering outperforms the feature-based clustering proposed by [10].

V. Conclusions

In this paper we address the problem of clustering for monaural blind source separation in the case of over-separation. A new iterative clustering algorithm is proposed. The results of the paper show that the iterative clustering improves the separation quality significantly for most of the mixtures and leads to more stable separation results than any other of the tested clustering algorithm. For a trade-off between separation quality and computational complexity the N-step clustering and the two-step clustering are introduced.

References

[1] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

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

[3] S. Dubnov. Extracting sound objects by independent subspace analysis. In *Proceedings of the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES22)*, June 2002.

[4] EBU. Sound Quality Assessment Material, Tech 3253, 1988. http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/.

[5] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 763–768, Nara, Japan, April 2003.

[6] H.G. Kim, Moreau N., and Sikora T. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[7] M. K. I. Molla and K. Hirose. Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):893–900, 2007.

[8] I. Orife. Riddim: A rhythm analysis and decomposition tool based on independent subspace analysis. Master's thesis, Dartmouth College, 2001.

[9] T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[10] J. Wellhausen. Audio signal separation using independent subspace analysis and improved subspace grouping. *Proceedings of the 7th Nordic Signal Processing Symposium*, pages 310–313, June 2006.