REAL TIME SEPARATION OF CONVOLUTIVE MIXTURES

Wolf Baumann, Bert-Uwe Köhler, Dorothea Kolossa and Reinhold Orglmeister

Berlin University of Technology
Institute of Electronics, Einsteinufer 17, 10587 Berlin
{w.baumann, b.koehler, d.kolossa, orglmeister}@ee.tu-berlin.de

ABSTRACT

A new concept of combining conventional beamforming with independent component analysis (ICA) techniques and its implementation on a multi DSP system is presented. The system consists of two floating point digital signal processors TMS320C6701, an eight channel linear microphone array, an analog/digital converter board and a handheld control unit for stand alone operation. In the two system stages a sum and delay beamformer as well as a convolutive ICA algorithm are implemented. Due to the high performance of the digital signal processors, the systems achieves blind separation of two convolutive mixed sources in real time.

1. INTRODUCTION

Recently, independent component analysis (ICA) has gained great importance in the field of blind source separation. Many algorithms are available for blind separation of instantaneously mixed signals, e.g. [1, 2, 3]. Applications have been reported e.g. in [4, 5].

In the instantaneous case the mixing process can be expressed in terms of weighted additions. This is not true for microphone recordings, where the mixing process yields convolutions with the room impulse responses between the sources and the microphones. Instead of the determination of scalar mixing weights, in the convolutive case impulse responses need to be identified.

A common approach to convolutive ICA is to transform the mixed signals into the frequency domain via short time Fourier Transform (STFT) and to solve the source separation problem within each of the frequency bins separately using an instantaneous ICA algorithm, e.g. [6, 7, 8]. The individual solutions of the ICA algorithms form a digital filter that can be applied either in the time or the frequency domain.

An important problem inherent in this technique are the permutations between the frequency bands. A subsequent rearrangement of the individual frequency bins on the basis of their similarity is a possibility to deal with this problem, but this method suffers from long room impulse responses.

Preprocessing by a beamformer can shorten the effective filter length by rejecting some of the reflections from the surrounding walls, resulting in a generally smoother frequency response. That is, a permutation correction criterion based on similarities between adjacent frequency bins may perform better with a beamforming preprocessing than without.

In order to analyse the combination of beamforming and convolutive ICA algorithms in real world environments we have implemented a system based on two digital signal processors (DSP) TMS320C6701 to perform beamforming and convolutive ICA in real time. First results of this system are presented here.

2. METHODS

This section gives a brief overview of the employed beamforming and ICA algorithm and discusses some issues concerning the realtime implementation.

2.1. Beamforming

The objective of the beamforming stage is the suppression of diffuse echos which appear in a typical room environment due to surrounding walls, floor and ceiling. These reverberations can be assumed to impinge mainly from outside the preferred look directions, which are adaptively aligned to the two sources of interest.

Regarding computational power and robustness, a delay and sum beamformer is a good choice. It can be easily applied in the frequency domain. This avoids the restriction to few look directions, as it could be caused by integer delay numbers after sampling.

The acoustical beamformer consists of two stages:

- estimating direction of arrival for two sources
- performing delay and sum beamforming for both look directions

2.1.1. Estimating direction of arrival

The implemented direction estimator is energy based. Due to possible source movements, it operates adaptively and has a sufficiently large buffer of mean directions to ensure robust estimation.

First, the complete area of interest is scanned in steps of five degrees. The two largest maxima are compared to two buffered mean directions. If the criterion of limited deviation is met, i.e. the estimated angles are plausible, the assignment is carried out.

In practice some additional constraints must be used for a successful direction estimation. So realizing a conventional delay and sum beamformer with an equidistant linear array, the beam pattern is strongly frequency dependent. That is, for low frequencies only negligible degrees of signal-to-noise ratio (SNR) improvement are achievable, while for higher frequencies with wavelengths λ above the half distance between the sensors, spatial aliasing may occur. Therefore, the direction finding is constrained to the optimal frequency f_{opt}, where the wavelength matches the sensor distance d, i.e. $f_{opt} = c/2d$ with c as the speed of sound propagation.

Additionally, the direction estimation is only performed in periods of high sensor activity, otherwise both look directions are held constant. This avoids adapting to wrong sources, like fan noise, if the source activity is temporarily low. For activity detection, the current signal power is compared to an adaptive threshold, which is determined by the multiplication of the buffered mean energy by a constant factor $\epsilon \approx 2.5$.

2.1.2. Delay and sum beamforming

The principle of delay and sum beamforming is the reinforcement of the signal impinging from the direction of interest, achieved by in phase summation of the sensor signals, see e.g. [9, 10]. In case of known source locations, the appropriate delays can be derived directly from the time it takes for the signal to propagate from one sensor to the next.

Operating in the frequency domain, the signals are multiplied by a frequency and look direction dependent phase shift factor Δ_k. The summation then can be expressed as

$$Y(\Omega, \tilde{t}) = \sum_{k=1}^{N} \alpha_k X_k(\Omega, \tilde{t}) \cdot e^{j\Omega\Delta_k}, \quad (1)$$

where $X_k(\tilde{t}, \Omega)$ is the short time fourier transformed signal of the k-th sensor and α_k is the window coefficient of the spatial sampling process.

Assuming two sources, this operation is performed twice on the same set of sensor signals for both look directions.

2.2. ICA

The instantaneous mixing process is defined as a linear combination of the statistically independent sources

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t), \quad (2)$$

where $\mathbf{s}(t), \mathbf{x}(t)$ and $\mathbf{A}$ denote the vector of the source signals, the vector of the mixed signals and the mixing matrix, respectively. The estimation of the matrix $\mathbf{A}$ or its inverse $\mathbf{W} = \mathbf{A}^{-1}$ is possible only up to scaling factors and permutations. The source signals are reconstructed by

$$\mathbf{u}(t) = \hat{\mathbf{W}} \cdot \mathbf{x}(t), \quad (3)$$

with $\hat{\mathbf{W}}$ as the estimated inverse mixing matrix. In the case of convolutive mixtures, the mixed signals not only contain scaled but also delayed versions of the source signals. For microphone recordings this is caused by the convolution of the source signals with the room impulse responses between source and sensor. It can be expressed as

$$\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t), \quad (4)$$

with $\mathbf{A}$ now containing the impulse responses of the mixing filters. Transformed into frequency domain, under certain circumstances the convolution can be reduced to a frequency dependent multiplication

$$\mathbf{X}(\Omega, \tilde{t}) = \mathbf{A}(\Omega) \cdot \mathbf{S}(\Omega, \tilde{t}), \quad (5)$$

where $\tilde{t}$ represents the index of the STFT frame. By applying a complex valued instantaneous ICA algorithm to every frequency band, the convolutive separation problem can be solved.

The elements of $\mathbf{A}(\Omega)$ form the frequency responses of the mixing filters. Unmixing is done with inverse filtering, i.e. with the elements of $\mathbf{W}(\Omega)$ in the frequency domain[1].

2.2.1. Jade

As an instantaneous ICA method, the JADE algorithm [1] is applied in each frequency band of the spectrogram to identify $\mathbf{A}(\Omega)$. This algorithm uses fourth order cumulants for estimating the unmixing matrix. It is based on the joint diagonalization of the most significant cumulant matrices.

First, the signals are decorrelated with a whitening matrix $\mathbf{M}$ (see e.g. [12])

$$\mathbf{X_s}(\Omega, \tilde{t}) = \mathbf{M}(\Omega)\mathbf{X}(\Omega, \tilde{t}) = \mathbf{M}(\Omega)\mathbf{A}(\Omega)\mathbf{S}(\Omega, \tilde{t}). \quad (6)$$

Hence, the following search is reduced to find an orthogonal rotation matrix $\mathbf{O}(\Omega)$ for separation. This is performed

[1] Time domain filtering is also possible by transforming back the filter coefficients, see [11] for an overview of blind separation for audio signals.

by making a set of cumulant matrices $Q_i^{X_s}$ as diagonal as possible, i.e. solving

$$\mathbf{O}(\Omega) = argmin \sum_i \text{Off}[\mathbf{O}^H(\Omega) Q_i^{X_s} \mathbf{O}(\Omega)]. \quad (7)$$

The matrix $\mathbf{O}(\Omega)$ is used to separate the mixed signals as follows

$$\mathbf{U}(\Omega, \tilde{t}) = \mathbf{O}^H(\Omega) \mathbf{X}_s(\Omega, \tilde{t}). \quad (8)$$

As for all ICA algorithms, a solution can be found up to uncertainties of scaling and permutation. In frequency domain based algorithms this leads to strong degradations in the separation result.

An additional problem arises from the blockwise execution of JADE in combination with real time processing. Because the unmixing filters are always computed anew, it is not sure how the output signals will be assigned to the channel numbers. Depending on the overall permutation, the channels may swap. So postprocessing is necessary to avoid the frequency and time domain permutations.

2.2.2. Scaling and frequency domain permutation

It can be shown [13] that, in a least squares sense, the optimal scaling factors correspond to the elements of the mixing matrix $\mathbf{A}$. So, multiplying the separated signals by coefficients obtained from the mixing matrix $\mathbf{A}$ yields compensation of scaling errors. This operation is equivalent to assuming a normalized mixing model with ones at the diagonal elements of $\mathbf{A}$. Although this supposition is not conform to the real mixing process, it has proven to be sufficient for a successful separation.

To correct permutations between frequency bands, it is assumed that the transfer function that results from concatenating all mixing matrices $\mathbf{A}(\Omega_k)$ will be smooth, which is true in case of a beamforming preprocessor. Thus, at each frequency, that permutation is selected, which leads to the smallest distance between the matrices $\mathbf{A}(\Omega_k)$ at the current band and $\mathbf{A}(\Omega_{k-1})$ at the previous band.

For this purpose, two matrices $\mathbf{A}^{(1)}(\Omega_k)$ and $\mathbf{A}^{(2)}(\Omega_k)$ are calculated, assuming the case of permutation and the case of no permutation, respectively. They are compared to the mixing matrix in the previous frequency band $\mathbf{A}(\Omega_{k-1})$, with the distance calculated according to

$$D_1 = \sum_{i,j} |a_{ij}^{(1)}(\Omega_k) - a_{ij}(\Omega_{k-1})| \quad (9)$$

$$D_2 = \sum_{i,j} |a_{ij}^{(2)}(\Omega_k) - a_{ij}(\Omega_{k-1})|, \quad (10)$$

and that $\mathbf{A}^{(i)}$ with the smaller distance value is selected to form $\mathbf{A}(\Omega_k)$.

2.2.3. Time domain permutation

To avoid channel swapping, the arrangement of the output signals must be kept consistent. For that purpose, the temporal structure of the separated frequency bands is exploited. Because consecutive blocks of STFT frames are built with an overlap of up to 75%, depending on the processor load, they cover mainly the same temporary events and thus show strong correlations.

Empirical examinations showed that it is sufficient to use a maximum number of 11 frequency bins. They are smoothed by a low pass filter operation and afterwards correlated. So we obtain 11 permutation indices p_i with values of either one or zero, which can be used for the decision on time permutation

$$P_{time} = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{11} p_i \geq 6, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

3. HARDWARE IMPLEMENTATION

An overview of the system is shown in Figure 1. The input/output stage consists of four stereo codecs CS4215 with a precision of 16 bit. They sample the signals from eight boundary condenser microphones (AKG® C400PC) at a sampling frequency f_s=11kHz and are also used for digital/analog conversion of the output signals.

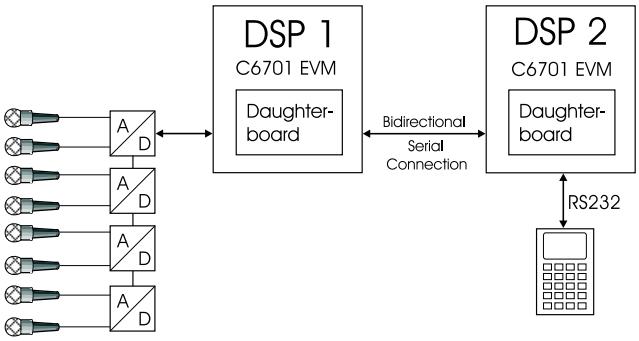

Fig. 1. Hardware structure

The high computational cost of ICA algorithms and the large amount of data in multichannel signal processing necessitates the use of two DSPs. We have used two TMS320C6701 EVM boards provided by Texas Instruments Inc. for our system and linked them via serial ports. This allows the bidirectional communication to run at 50 MHz.

The peripheral equipment, i.e. the codecs and the handheld control unit, are also connected via serial interfaces. The system is capable of stand alone operation, since the daughter boards are equipped with boot EEPROM. Changes of parameters are carried out by the microcontroller based handheld control unit.

The two DSPs run at a clock rate of 100 MHz and perform the individual tasks as follows:

DSP 1	◦ codec communication ◦ short time fourier transformation ◦ direction estimation ◦ sum & delay beamforming
DSP 2	◦ performing JADE for every frequency bin ◦ correction of frequency permutation ◦ correction of time permutation ◦ transform data back to the time domain

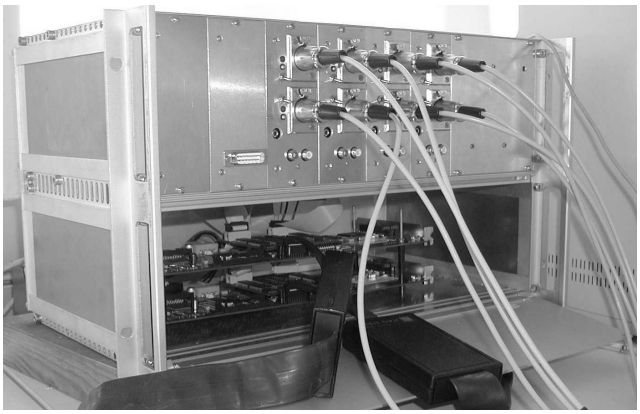

Fig. 2. Image of the compact system

4. EXPERIMENTAL SETUP AND RESULTS

The default parameters of the system are initialzed after startup and are set as shown in table 1.

Table 1. Default parameters

Parameter	Value
sampling frequency	11025Hz
number of microphones	8
spatial window function	hamming
DFT length	512
STFT hamming window length	256
STFT overlap	192
blocklength for JADE	1000
energy threshold factor ϵ	2.5

The recordings were taken in a normal office environment, i.e. a room with the dimensions 6m × 4.8m. Two loudspeakers were placed in a distance of about 1.7m at the angles shown in Figure 3. Previously made recordings of a male and a female voice served as sound sources.

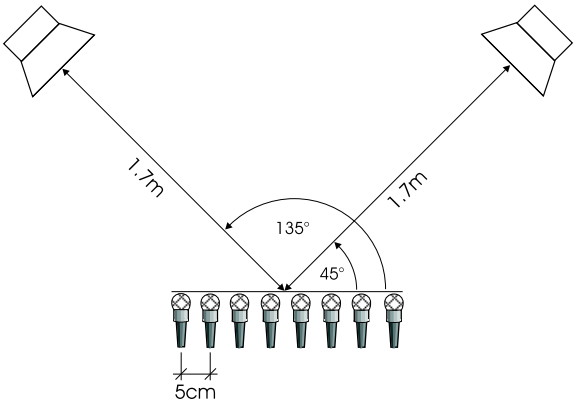

Fig. 3. Experimental setup

A common method to evaluate the quality of separation is to compute the signal to noise ratio (SNR) or its improvement. However, direct SNR estimation is difficult because of additional convolutions and overall scaling factors introduced by the ICA stage. To exclude errors in the SNR estimation, the source signals and the separated signals have to be made comparable, i.e. differences in variance and temporal delays have to be compensated.

Hence, we first define two desired signals, i.e. the best solutions to the separation problem, by alternately recording the sources $s_1(t)$ and $s_2(t)$. In the second step an appropriate scaling and time delay is applied. That is, the sources and the separated signals are normalized to unit variance

$$\tilde{s}_i(t) = \tfrac{1}{\sigma_{s_i}} \cdot s_i(t) \tag{12}$$

$$\tilde{u}_i(t) = \tfrac{1}{\sigma_{u_i}} \cdot u_i(t), \tag{13}$$

and the cross correlation between $u_i(t-\tau)$ and $s_i(t)$ is maximized so that the signals do not show significant timeshifts. The normalization to unit variance results in a zero dB SNR of the mixed signals, so that the SNR of the separated signals corresponds directly to the value of improvement.

For the calculation of the SNR of the separated signals an error signal is necessary. It can be estimated as the difference of the absolute values of the spectrograms of $\tilde{s}_i(t)$ and $\tilde{u}_i(t)$

$$E_i(\Omega, \tilde{t}) = |\tilde{S}_i(\Omega, \tilde{t})| - |\tilde{U}_i(\Omega, \tilde{t})|. \tag{14}$$

Eventually, the SNR of the separated signals is calculated in the frequency domain by

$$SNR_i = 10 \cdot \log \frac{\sum_\Omega \sigma^2_{S_i}}{\sum_\Omega \sigma^2_{E_i}}. \tag{15}$$

The SNRs of the separated speech signals are shown in the following table.

Table 2. Results expressed as SNR improvement

	mixed sources	improvement
$\frac{\text{male}}{\text{female}}$	0 dB	7.0 dB
$\frac{\text{female}}{\text{male}}$	0 dB	1.6 dB

5. DISCUSSION AND FUTURE WORK

It has been shown that the combination of beamforming and independent component analysis is well suited for the separation of real world signals in real time.

Although the current implementation has been proven to work in a real environment, enhancements are possible and necessary. From the computational point of view, the application of a batch algorithm, such as JADE, is a major drawback because of possible delays as well as possible permutations in time and frequency domain. Additionally, continuously varying recording conditions can only be handled by overlapping the blocks.

To enhance the quality of the separated signals several improvements of the algorithms are possible. Besides an additional denoising stage, a critical factor is the correct estimation of the directions of interest within the beamforming stage. The implementation of an eigenvalue based algorithm should improve the detection [14]. Another important way of improvement are frequency invariant beamformers [15].

The further work also includes the ongoing search for better permutation correction methods as well as new means for better cost functions. The algorithm proposed in [7] seems to be promising in this context.

6. REFERENCES

[1] J.-F. Cardoso and A. Souloumiac. Blind beamforming for Non Gaussian Signals. In *IEEE-Proceedings-F*, volume 140, pages 362–370. IEEE, dec 1993. http://sig.enst.fr/ cardoso/.

[2] A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.

[3] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9:1483–1492, 1997.

[4] S. Makeig, T.-P. Jung, D. Ghahremani, A. Bell, and T.J. Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proc. Natl. Acad. Sci USA*, 94:10979–10984, 1997.

[5] R. Vigario, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of eeg and meg recordings. *IEEE Trans. on Biomed. Eng.*, 47:589–593, 2000.

[6] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *BSIS Tech. Report, http://www.bsis.brain.riken.go.jp/*, 1998.

[7] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. In *IEEE Trans. on Speech and Audio Processing*, pages 320–327, May 2000.

[8] T.-W. Lee, A. Ziehe, R. Orglmeister, and T.J. Sejnowski. Combining timedelayed decorrelation and ica: Towards solving the cocktail party problem. *Proc. ICASSP*, 2:1249–1252, 1998.

[9] Martin Drews. *Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestörter Sprache*. PhD dissertation, Technische Universität Berlin, 2000.

[10] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, Englewood Cliffs, 1993.

[11] Kari Torkkola. Blind Separation for Audio Signals - Are We There Yet? In *ICA1999*, 1999. http://members.home.net/torkkola.

[12] Bert-Uwe Köhler. *Realzeitfähige blinde Quellentrennung am Beispiel elektroenzephalographischer Signale*. PhD thesis, Technische Universität Berlin, 1999.

[13] Wolf Baumann. Implementierung und Analyse von Verfahren der Blinden Quellentrennung für Schallsignalaufnahmen in realen Umgebungen. Master's thesis, Technische Universität Berlin, 1999.

[14] David K. Campbell. Adaptive Beamforming Using a Microphone Array for Hands-Free Telephony. Master's thesis, Virginia Polytechnic Institute and State University, 1999.

[15] Darren B. Ward, Rodney A. Kennedy, and Robert C. Williamson. Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. In *Journal of the Acoustical Society of America*. Acoustical Society of America, Sep 1994.