ON-LINE BLIND SEPARATION OF MOVING SOUND SOURCES

Jörn Anemüller and Tino Gramss (†)

Dept. of Physics, Graduate School in Psychoacoustics
Carl-von-Ossietzky-University, 26111 Oldenburg, Germany
ane@uni-oldenburg.de

ABSTRACT

In this paper, we propose a method for the on-line blind separation of sound sources in the case where the mixing filters have a δ-shaped impulse response. Our algorithm works entirely in the frequency domain and exhibits fast convergence due to cross-frequency couplings.

Specific problems related to on-line separation of running speech are discussed. The algorithm performs successful separation of digitally mixed speech signals and of signals from a moving and a standing speaker recorded in an anechoic chamber.

1. SOURCE SEPARATION

In the case of an instantaneous linear superposition

$$m_i(t) \quad = \quad \sum\nolimits_j A_{ij} s_j(t), \tag{1}$$

of N independent source signals $s_j(t)$ resulting in N sensor signals $m_i(t)$, various algorithms for blind source separation have been proposed; refer to, e.g., [4]. Their common goal is to find an estimate $\mathbf{W} = [W_{ij}]$ for the inverse of the mixing matrix $\mathbf{A} = [A_{ij}]$ and reconstruct the sources as $\vec{x}(t) = \mathbf{W}\vec{m}(t)$. (We use the notation $\vec{x}(t) = [x_1(t), \ldots, x_N(t)]^T$, etc.)

The source separation algorithm used by us is derived through the maximum-likelihood method [3, 6], augmented for complex-valued signals. Since we are dealing with Fourier transformations of speech signals, we model the probability densities of the corresponding spectral components s_j as

$$P(s_j) = P(||s_j||) \propto \cosh^{-1}(||s_j||) \tag{2}$$

where $||s_j||$ denotes the magnitude of the complex number s_j. Hence, s_j is zero-mean, all values of the complex phase are equally probable, and the distribution has positive kurtosis. This leads to the following update equation

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \eta \left(\mathbf{W}^{(t)} - \vec{u}(t)\vec{x}^H(t)\mathbf{W}^{(t)} \right), \tag{3}$$

$$\text{where} \quad u_i(t) = \frac{x_i(t)}{||x_i(t)||} \tanh(||x_i(t)||), \tag{4}$$

for the weight-matrix $\mathbf{W}$ at time-step t. The transposition and complex conjugation of $\vec{x}$ is denoted as $\vec{x}^H$. We denote by η the adaptation step's size.

In the remainder of this paper, two well-known symmetries of source separation algorithms are relevant: scaling-invariance, i.e., the sources can only be reconstructed up to a constant scaling factor, and permutation-invariance, i.e., it is unknown which reconstructed signal resembles which source.

2. ACOUSTIC SOURCE SEPARATION IN THE FREQUENCY DOMAIN

When mixing acoustic sources, the assumption of an instantaneous superposition does not hold, since propagation delays and echoes have to be taken into account. Accordingly, multiplication in Eq. 1 is replaced by the convolution of the sources signal $s_j(t)$ and the impulse response $A_{ij}(t)$ of the room from source j to microphone i:

$$m_i(t) \quad = \quad \sum\nolimits_j \int_\tau A_{ij}(\tau) s_j(t-\tau). \tag{5}$$

The filtering with the room-transfer function can be expressed more elegantly in the frequency domain [1, 2]. By computing short-time spectra $\hat{m}_i(f,T)$ and $\hat{s}_i(f,T)$ at times $T = 0, \Delta T, 2\Delta T, \ldots$ of $m_i(t)$ and $s_i(t)$, respectively, the discrete approximation of Eq. 5 is obtained as

$$\hat{m}_i(f,T) = \sum\nolimits_j \hat{A}_{ij}(f)\hat{s}_j(f,T), \quad f = f_1, \ldots, f_M. \tag{6}$$

Provided the impulse response $A_{ij}(t)$ is much shorter than the time-window used for computing the short-time spectra, this approximation is sufficiently accurate for practical purposes.

In Eq. 6 we have transformed the acoustic source separation problem into M independent instantaneous source separation problems, one for each frequency $f_1, \ldots, f_M$. Each subproblem can be solved separately with the algorithm Eq. 3. However, permutation and scaling of the reconstructed signals with respect to the sources will in general be different for each frequency f_i. Without further precautions, this hinders reconstruction of the source signals in the time-domain.

(†) Dr. Tino Gramss, who initiated and supervised this work, died in January 1998.

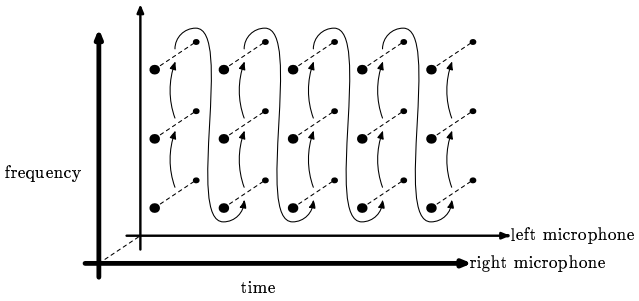

Figure 1: Schema showing the order of scanning the spectrograms of the microphone signals. Each dot represents a spectral component in a spectrogram.

Furthermore, source separation is achieved relatively slow due to the low sampling rate $1/\Delta T$ of the spectrograms.

3. SOURCE SEPARATION FOR A δ-SHAPED IMPULSE RESPONSE.

We consider source separation in rooms with a δ-shaped impulse response, i.e., where propagation-time- and level-differences do occur, but echoes do not. This matches sound propagation in an ideal anechoic chamber. In this case, the superposition can be expressed in the time-domain as[1]

$$m_i(t) \quad = \quad \sum\nolimits_j A_{ij} s_j(t - \tau_{ij}), \quad A_{ij} > 0 \text{ real.} \quad (7)$$

After a Fourier transform, the propagation time τ_{ij} from source j to microphone i results in a rotation of the complex phase $\arg \hat{A}_{ij}(f) = -2\pi f \tau_{ij}$ of the mixing coefficients $\hat{A}_{ij}(f)$, while their magnitude $||\hat{A}_{ij}(f)|| = A_{ij}$ remains constant:

$$\hat{m}_i(f,T) \quad = \quad \sum\nolimits_j \hat{A}_{ij}(f) \hat{s}_j(f,T), \quad (8)$$

$$\text{with} \qquad \hat{A}_{ij}(f) \quad = \quad A_{ij} e^{-2\pi i f \tau_{ij}}. \quad (9)$$

This relationship among the values of the transfer function at different frequencies makes possible to iterate the separation algorithm across frequencies: When scanning the spectrograms of the recorded signals according to Fig. 1, the information obtained at low frequencies resolves the ambiguity between τ_{ij} and $\arg \hat{A}_{ij}(f)$, whereas high frequencies improve on the accuracy of the estimate for τ_{ij}.

In detail, our algorithm consists of the following steps:
1) Compute the stochastic gradient $\Delta W_{ij}(f)$ for the current frequency f, using the standard source separation algorithm (Eq. 3). The improved estimate for $[\hat{A}_{ij}(f)]$ is given by $[\hat{A}^{(e)}_{ij}(f)] \equiv [W_{ij}(f)]^{-1}$.

[1]The assumption $A_{ij} > 0$, which is true for acoustic sources, can easily be removed.

2) Compute the corresponding stochastic gradients for time and level differences as

$$\Delta\tau^{(e)}_{ij} \quad = \quad -\Delta \left[\arg \hat{A}^{(e)}_{ij}(f) \right] / (2\pi f), \quad (10)$$

$$\Delta A^{(e)}_{ij} \quad = \quad \Delta \left[||\hat{A}^{(e)}_{ij}(f)|| \right], \quad (11)$$

respectively, where the superscript (e) denotes the current estimates of the corresponding quantities. (It is beneficial to replace Eq. 11 by Eq. 12; refer to section 4.)
3) Switch to the next higher frequency. If the highest frequency has been iterated, switch to the lowest frequency of the next short-time-spectrum (Fig. 1).

This procedure solves the aforementioned problems of source separation in the frequency domain: By 'linking' the different frequencies with each other, permutation of the output-channels of the algorithm at different frequencies are avoided. Furthermore, the scaling problem is avoided since only temporal and level differences between microphones are important in signals which involve delays and level differences only. Therefore, a time-domain reconstruction of the source signals becomes possible. Finally, fast convergence of the algorithm is achieved by its effective use of all spectral components for parameter estimation.

4. ON-LINE SEPARATION OF RUNNING SPEECH

On-line separation of continuous speech, as opposed to batch separation of a given chunk of speech, poses additional problems. These are mainly due to the time-varying statistics of speech signals, when analyzed on a time-scale of seconds.

Since the short-time spectra are computed from 30ms batches of the microphone signals, this suggests estimating update-steps for the separating matrix in a batch-like manner from each spectrum. However, a batch-algorithm cannot be employed for this task, since our algorithm relies on switching serially from low to high frequency components. In order to still obtain an accurate estimate of the separating matrix, we use the effective step-size $\eta_{\text{eff}}(f) = \eta/f$ for frequency f, which gives equal weight to all spectral components [3]. As a result of Eq. 9, this step-size is already 'built-in' to the adaptation of $\tau^{(e)}_{ij}$ (Eq. 10). Replacing Eq. 11 by

$$\Delta A^{(e)}_{ij} \quad = \quad \Delta \left[||\hat{A}^{(e)}_{ij}(f)|| \right] / f, \quad (12)$$

we use $\eta_{\text{eff}}(f)$ for the adaptation of $A^{(e)}_{ij}$ as well. We have found that this improves on stability and accuracy of the algorithm.

Another problem are speech pauses in one source which, in the examples of section 5, last up to 700 milliseconds. Without additional precautions, the algorithm would

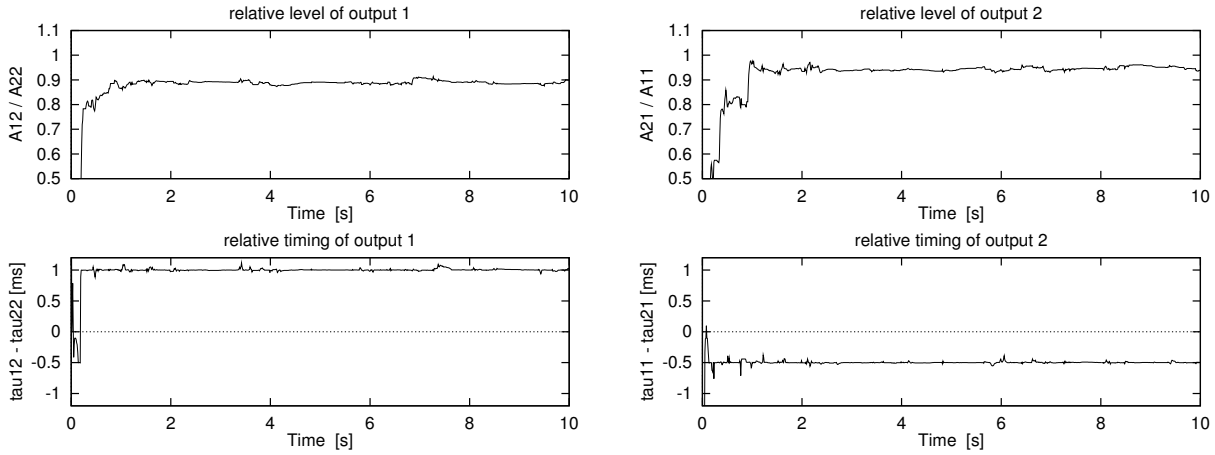

Figure 2: Time-course of relative level and timing for the separation of artificially mixed sources.

diverge during these intervals, i.e., it would attempt to find an alternative source to be separated. We prevent this by introducing a relative threshold for the power of the sources: If the energy of any reconstructed signal in the current FFT-frame is less than 15% of the energy of the other reconstructed signal, then solely separation but no parameter update is performed.

5. EXPERIMENTS

We have performed experiments with artificially mixed sources and with real-world recordings in an anechoic chamber. In this paper we present two experiments: In the first one, we verify the proposed algorithm using speech signals which have been mixed digitally in the time-domain. In the second experiment, source separation is performed on real-world recordings made in an anechoic chamber. Here, one of the sources is moving while the second is stationary.

In both cases the following preprocessing was used in order to obtain the input spectrograms: The signals were recorded using a sampling rate of 48 kHz. Speech pauses were not removed. A preemphasis $(x(t+1) - x(t))$ was used in order to eliminate low-frequency noise and increase the energy of the high-frequency components of the speech signals. Spectrograms were computed using a Hanning-window of length 30 ms and a window-shift of 10 ms. The resulting frames were padded with zeros to 2048 samples before a Fast-Fourier-Transform was applied. Spectral components from 23Hz to 10kHz were used for adaptation, since the main energy of the signals occurs in this range.

The parameters of the algorithm were initialized to $A_{ij} = \delta_{ij}$ and $\tau_{ij} = 0$, i.e., the algorithm started off from the (wrong) assumption that no mixing occurs. This imposed the additional difficulty of breaking the symmetry of the initial state. The initial learning rate was set to $\eta = 0.4$ in order to pass first transients. It was then lowered proportionally to $1/T$ until it reached $\eta = 0.001$ after 4 seconds. $\eta = 0.001$ was then kept constant for the remaining time.

Finally, the separated signals were transformed back to the time-domain, using the overlap-add method [5].

5.1. Artificially mixed sources

Two mono speech signals were digitally 'stereofied' and mixed in the time-domain, using relative timing and level of $\tau_{21} - \tau_{11} = 0.5$ ms and $A_{21}/A_{11} = 0.95$, respectively, for the first source, and $\tau_{12} - \tau_{22} = 1.0$ ms and $A_{12}/A_{22} = 0.90$, respectively, for the second source.

Fig. 2 shows the time-course of the relative timing and level for both reconstructed signals. It is clearly visible, that both sources are reconstructed approximately correct after 1 second. However, due to the non-stationary nature of speech signals, the parameters fluctuate in time.

When listening to the reconstructed signals, an almost inaudible crosstalk, together with soft 'musical noise' artifacts, can be heard.

The fast and almost perfect separation demonstrates that the proposed algorithm operates successfully under optimal conditions.

5.2. Moving sources in an anechoic chamber

In the second example, speech signals from a moving and a stationary speaker are separated. The stereo recording was performed in the anechoic chamber of the University of Oldenburg.

The experimental setup was as follows: Two microphones were placed 35 cm apart. The stationary speaker was standing in a distance of 3 m at 60 degrees left of the mid-perpendicular of the microphones. The moving speaker started at a distance of 4 m at 70 degrees to the right. He

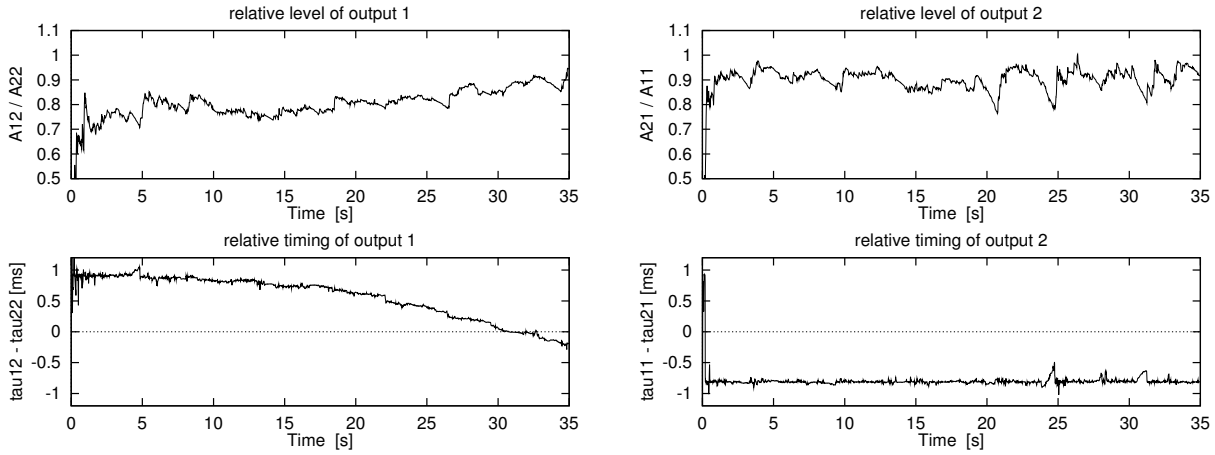

Figure 3: Time-course of relative level and timing for the separation of a moving and a stationary speaker recorded in an anechoic chamber.

walked in a straight line parallel to the microphones until he reached a position at about 30 degrees left.

Preprocessing and learning-parameters were the same as in the first example.

The time-course of the relative timing and level for both reconstructed signals is shown in Fig. 3. Again, the algorithm converges within about 1 second. The movement in the second speaker signal is clearly visible from the slow drift in level and timing parameters of the first output channel. While the estimates for the relative timing are accurate and exhibit only small short-term fluctuations, the estimates for the relative level show larger fluctuations.

Listening to the reconstructed signals, the crosstalk is still soft, but larger than in the first example. Also, additional noise can be heard as a side-effect of the separation.

This is a result of the real-world situation which introduces deviations from a perfectly δ-shaped impulse response. In particular, even a good anechoic chamber produces small reflections, and any two microphones differ slightly in their characteristics.

6. CONCLUSION

We have proposed an algorithm for the separation of mixtures involving time-delays and level-differences, i.e., a δ-shaped impulse response of the mixing filters. Separation is performed entirely in the frequency-domain. Fast convergence of the algorithm is achieved by exploiting cross-frequency couplings and switching sequentially between frequencies. The algorithm has been applied to artificial and real-world mixtures of running speech.

Problems related to on-line separation of running speech have been addressed using step-size adaptation and speech pause detection.

Experiments with artificially mixed speech signals have shown convergence within 1 second towards almost perfect signal separation.

Experiments with signals from a standing and a moving speaker, recorded in an anechoic chamber, have shown a similar speed of convergence. However, the quality of separation is not as high as in the previous experiment.

Directions for further research include automatic compensation of slight deviations from δ-shaped impulse responses and incorporation of a-priori knowledge about, e.g., head related transfer functions of dummy heads.

7. REFERENCES

[1] V. Capdevielle, C. Servière, and J. L. Lacoume. Blind separation of wide band sources in the frequency domain. In *ICASSP 1995*, pages 2080–2083, 1995.

[2] T. Gramss. A neural model for the separation of acoustic signals. In J. Bower, editor, *Computational Neuroscience: Trends in Research 1995*, pages 191–195, Monterey, July 1995.

[3] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis, draft 3.7. URL: ftp://wol.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz, Dec. 1996.

[4] J.-P. Nadal and N. Parga. Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Computation*, 9:1421–1456, 1997.

[5] A. V. Oppenheim and R. W. Schaefer. *Digital signal processing*. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs, NJ, 07632, 1989.

[6] D. T. Pham, P. Garat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications*, pages 771–774, 1992.