Convolutive Blind Source Separation of Speech Signals

Jörn Anemüller* and Birger Kollmeier
Medizinische Physik, Universität Oldenburg

*presently: The Salk Institute for Biological Studies and
Swartz Center for Computational Neuroscience, San Diego

Blind Source Separation

The problem of blind source separation (BSS) is encountered in various applications where it is desired to reconstruct multiple original source signals while only mixtures of them can be observed. Lack of additional information, e.g., about spatial locations of the sources, is indicated by the term ‘blind’. One example is the area of noise reduction algorithms where the aim is to separate out a speech signal from a background of noise or competing speech signals, in order to enhance speech intelligibility for hearing aid users or to improve the recognition rate of automatic speech recognition systems. Many further applications exist in domains such as image processing, biomedical data analysis and document analysis. To separate convolutively mixed source signals, filtering of the microphone signals must be performed—instead of a multiplication in the case of non-convolutive mixing. Depending on the domain in which the filters are implemented, algorithms from the literature fall into the classes of time domain or frequency domain based algorithms.

We here describe the amplitude modulation decorrelation (AMDecor) algorithm for convolutive blind source separation, that separates the signals in the frequency domain. Details of the algorithm and its evaluation are covered in [Ane01]. Frequency domain algorithms are based on the property of the Fourier transformation that the convolution in the time domain results in a multiplication in the frequency domain. Thereby, the convolutive source separation problem in the time domain is transformed into K decoupled instantaneous source separation problems in the frequency domain, one for each frequency $f = 1, \ldots, K$ (e.g., [AK03]). After separation has been performed in the frequency domain, the separated sources are transformed back to time domain signals using, e.g., the overlap-add technique. The drawback of frequency domain methods is that in general local permutations may arise, i.e., the sources’ spectral components are recovered in a different (unknown) order in different frequency channels, thereby making a time domain reconstruction of the source signals impossible.

The AMDecor Algorithm

The AMDecor algorithm described here employs a novel cost-function that integrates information across different frequencies in order to perform separation. Unlike existing methods, the proposed algorithm employs correlations of signal envelopes at different frequencies. It is shown that this approach solves the problem of local permutations and results in a good quality of signal separation.

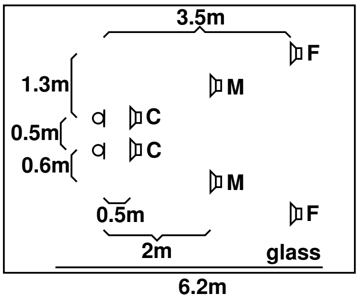

Abbildung 1: The setup for the real-room recordings with microphones to the left and speakers at close (C), medium (M) and far (F) positions to the right. Direct to reverberation ratio (DRR) for the three conditions was 4.7 dB (C), -1.0 dB (M), and -6.9 dB (F), respectively.

The basis for the AMDecor algorithm is formed by the highly interrelated amplitude modulation in different and even distant frequency channels, that is a property of speech signals. This property is termed *amplitude modulation (AM) correlation*. A natural way to measure the synchrony of the amplitude modulation in two frequency channels of two (possibly different) signals is to compute the correlation between the corresponding frequency specific signals envelopes. Due to the low-pass filtering property of the magnitude operation, the envelope correlation can be computed as the correlation of the time-courses in two frequency channels of amplitude spectrograms. The amplitude modulation correlation (*AMCor*) $c(x(T, f_k), y(T, f_l))$ between the frequency channel f_k of spectrogram $x(T, f)$ and frequency channel f_l of spectrogram $y(T, f)$ is defined as

$$\begin{aligned} c(x(T, f_k), y(T, f_l)) &= E\{|x(T, f_k)|\,|y(T, f_l)|\} \\ &- E\{|x(T, f_k)|\}\,E\{|y(T, f_l)|\}. \end{aligned}$$

Since amplitude modulation correlation provides a measure of the similarity of speech signals, it can be used as a criterion for blind source separation by requiring that the AM correlation for any two frequency bands of all different reconstructed signals vanishes.

This requirement of amplitude modulation *de*correlation of the unmixed signals solves both the blind source separation problem and the problem of local permutations simultaneously. Figure 3 illustrates the processing schema underlying the AMDecor algorithm.

Evaluation

The capabilites of the AMDecor algorithm were evaluated in several experiments [Ane01]. Synthetic source signals were used to show that the algorithm successfully separates signals which are inseparable for algorithms working in isolated frequency channels.

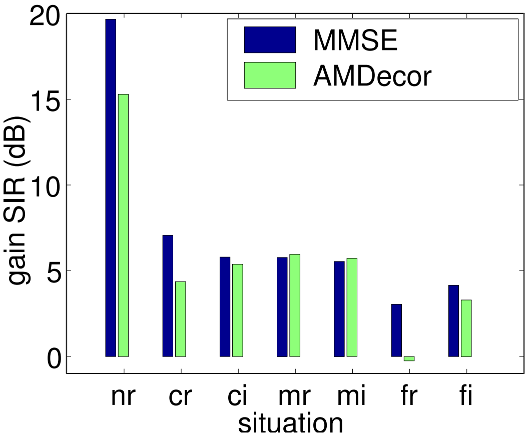

Abbildung 2: Comparison of the gain in SIR accomplished by AM decorrelation algorithm (‘AMDecor’) and MMSE method (‘MMSE’). The different acoustic situations are non-reverberant (‘nr’), close (‘cr’, ‘ci’), medium (‘mr’, ‘mi’) and far (‘fr’, ‘fi’). Data for ‘nr’, ‘cr’, ‘mr’ and ‘fr’ was obtained by sound recordings in a room, while data for ‘ci’, ‘mi’ and ‘fi’ was obtained by convolving the original source signals with impulse responses measured in the room.

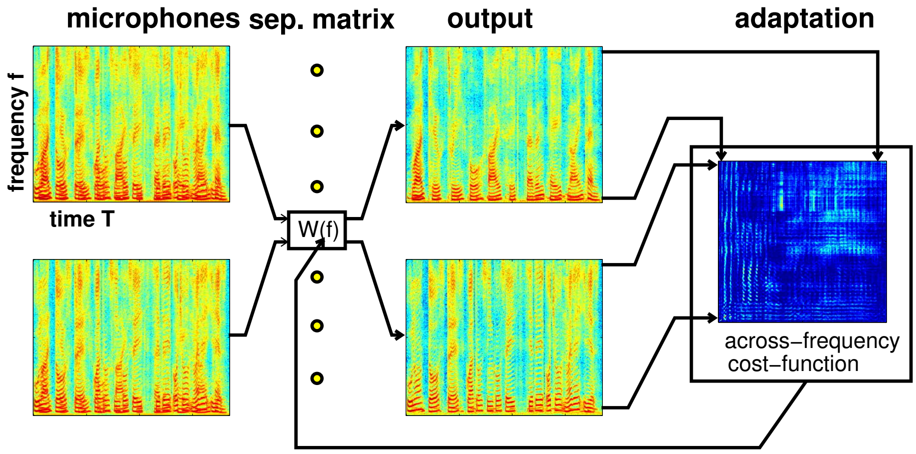

Abbildung 3: Schematic diagram of the AMDecor algorithm. From left to right: The spectrograms of the recorded signals are transformed by frequency-dependent matrices $\mathbf{W}(f)$, resulting in unmixed signals. The adaptation loop of the AMDecor algorithm adapts the $\mathbf{W}(f)$ until the unmixed signals exhibit minimal amplitude modulation correlation.

Quality of separation and ability to avoid local permutations were addressed using real-room recordings from several acoustic situations. Performance on publicly available benchmark datasets was compared with results from previous blind source separation algorithms. Demonstration sound files can be obtained from `http://medi.uni-oldenburg.de/demo/ane/diss`.

We here present some results obtained in different acoustic situations which include recordings in reverberant and non-reverberant environment and data obtained by digitally convolving speech signals with impulse responses of a real room. Since the recordings were performed such that the original source signals are available, and since the corresponding room impulse responses have also been recorded, the recordings allow for a detailed evaluation and analysis of the AM decorrelation algorithm. In particular, the results of the AMDecor algorithm are compared to the (non-blind) MMSE solution.

Speech signals from a total of four acoustic situations were recorded in two rooms. In the first room, a medium-sized seminar room at University of Oldenburg, speech was recorded from three different distances between speakers and microphones, ranging from 0.5 m to 3.5 m. The details of the setup are shown in figure 1. Note that the room contained a large window front to the right of the microphones and a blackboard behind the microphones. These surfaces contribute very strong reflections to the room acoustics, resulting in a reverberation time T_{60} of 0.5 s. In addition, recordings were also performed in the anechoic chamber of the University of Oldenburg.

Figure 2 compares the separation results obtained by the AM decorrelation algorithm with separation by the MMSE method. In the case of simulated mixing, the result of the AMDecor separation is close to the MMSE result and in one case even better. In the case of real recordings, AMDecor performs on average slightly worse, however still close to the MMSE result. In the ‘medium’ situation, AMDecor outperforms MMSE. In the problematic case of the ‘far’ situation, the AMDecor algorithm converges to a local optimum with overall poor separation, that is a result of local permutations (see [Ane01]). In conclusion, the separation obtained by AM decorrelation is on average in the vicinity of the optimum.

We also applied the AM decorrelation algorithm to publicly available real room recordings of speech signals. The quality of separation is evaluated and compared to the quality accomplished by previous algorithms on the same data. Since the original source signals are not available, we resorted to use the value of the AMDecor cost-function as an indication for the algorithms' performance and used listening tests to assess subjective quality of separation. The results are displayed in figure 4. They show that the AM decorrelation algorithm has the best performance and improves slightly on the algorithm of [PS00] which is generally regarded as exhibiting high-quality output signals and excellent separation results. Informal listening tests confirmed that perceived quality of separation is in good agreement with the obtained numerical values.

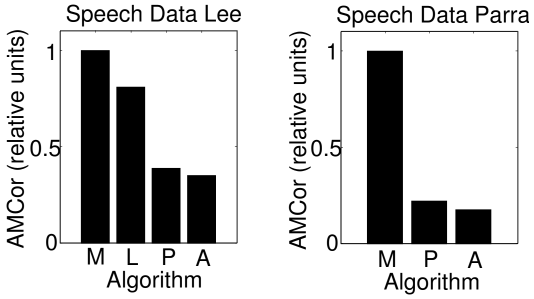

Abbildung 4: Separation of real-room recordings of speech by different algorithms. Left side refers to speech data recorded by [Lee98], right side to speech data recorded by [Par98]. The value of the AMCor cost-function (‘AMCor’) is shown for the amplitude modulation decorrelation algorithm (‘A’), Lee's algorithm (‘L’, [LZOS98], results for Lee's speech data only) and Parra's algorithm (‘P’, [PS00]), relative to the mixed signals (‘M’). Total energy of the signals was normalized prior to computing the cumulative AMCor.

This work was supported by the German Research Council (DFG).

[AK03] J. Anemüller and B. Kollmeier. Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach. *Speech Communication*, 39(1-2):79–95, January 2003.

[Ane01] J. Anemüller. *Across-Frequency Processing in Convolutive Blind Source Separation*. PhD thesis, Dept. of Physics, University of Oldenburg, Oldenburg, Germany, 2001. `http://medi.uni-oldenburg.de/members/ane`.

[Lee98] Te-Won Lee. Sound recordings `rss_mA.wav` and `rss_mB.wav`, 1998. `http://tesla-e0.salk.edu/~tewon/Blind/Demos`.

[LZOS98] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. J. Sejnowski. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1249–1252, Seattle, USA, May 1998.

[Par98] Lucas Parra. Sound recordings `tvin1.wav` and `tvin2.wav`, 1998. `http://www.sarnoff.com/career_move/tech_papers/papers`.

[PS00] Lucas Parra and Clay Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8:320–327, 2000.