To appear in: Joint meeting of the Acoustical Society of America and the European Acoustics Association, Berlin, Germany, March 14-19, 1999.

Correlated modulation: a criterion for blind source separation

Jörn Anemüller

Medical Physics Group and Graduate School in Psychoacoustics
Carl von Ossietzky-University, D-26111 Oldenburg, Germany
e-mail: ane@uni-oldenburg.de

Summary: The problem of blindly separating a convolutive mixture of modulated signals is considered. Spectrograms of the signals are computed and separation is performed in the frequency domain. A new algorithm for blind source separation is proposed, which is based on correlated modulation in the sources' different frequency channels. For example, speech contains correlated modulation in different frequency regions. The algorithm successfully separates mixtures of modulated artificial signals and of speech.

INTRODUCTION

The goal of blind source separation is to reconstruct mutually independent source signals $s_1(t), \ldots, s_N(t)$ when only mixtures $m_1(t), \ldots, m_M(t)$ of them can be observed. A minimum set of assumptions regarding the probability density functions or the autocorrelation functions of the sources are made in order to accomplish this. In particular, the geometry of the sensors and sources is completely unknown, hence the term 'blind'.
The case of an instantaneous, linear, square and invertible mixing of the sources is well-understood. Here, the mixtures are given by $\mathbf{m}(t) = [m_1(t), \ldots, m_N(t)]^T = \mathbf{A}\mathbf{s}(t) = \mathbf{A}[s_1(t), \ldots, s_N(t)]^T$ where $\mathbf{A} = [a_{ij}]$ is the invertible $N \times N$ mixing matrix. Various algorithms, based on higher-order statistics or time-delayed correlations, exist for finding an estimate $\mathbf{W}$ of $\mathbf{A}^{-1}$ [2]. The sources can then be reconstructed as $\mathbf{x}(t) = [x_1, \ldots, x_N(t)]^T = \mathbf{W}\mathbf{m}(t)$. However, it is a principle limitation that $\mathbf{x}(t)$ can resemble $\mathbf{s}(t)$ only up to an unknown permutation and rescaling of the elements of $\mathbf{s}(t)$.
The situation is more involved for the acoustic superposition of sound sources, since the room´s impulse response gives rise to a convolutive mixing process. One may resort to the frequency domain by approximating the linear convolution in the acoustic medium by the circular convolution of the discrete Fourier transformation. By computing short-time spectra $\hat{\mathbf{m}}_f(T)$ and $\hat{\mathbf{s}}_f(T)$ of $\mathbf{m}(t)$ and $\mathbf{s}(t)$, respectively, at times $T = 0, \Delta T, 2\Delta T, \ldots$, the mixtures can be written as

$$\hat{\mathbf{m}}_f(T) = \hat{\mathbf{A}}_f \hat{\mathbf{s}}_f(T), \qquad f = 1, \ldots, \nu. \tag{1}$$

Hence, the acoustic source separation problem is transformed into ν independent instantaneous source separation problems, one for each frequency channel $f = 1, \ldots, \nu$.
The main problem at this point stems from the aforementioned indeterminacy with respect to permutation of the sources: one may find spectral components belonging to source i in the jth component of the reconstructed signals $\hat{\mathbf{x}}_\alpha(T)$ for frequency channel α, but in the kth component of $\hat{\mathbf{x}}_\beta(T)$ for frequency channel β, with $j \neq k$. Since this permutation is a-priori unknown, any time-domain reconstruction of the sources is impossible. Algorithms have been developed which

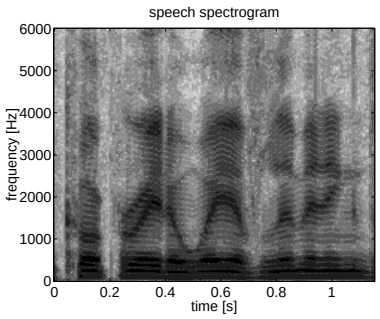

Figure 1: Spectrogram of speech containing correlated modulation in different frequency regions.

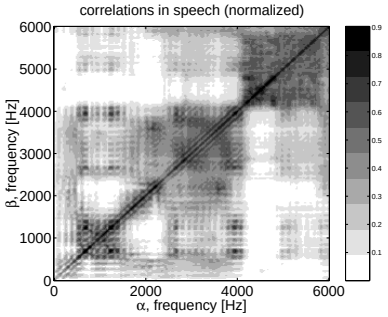

Figure 2: Example of correlation contained in a speech signal: $C(\hat{s}_{i,\alpha}, \hat{s}_{i,\beta})$.

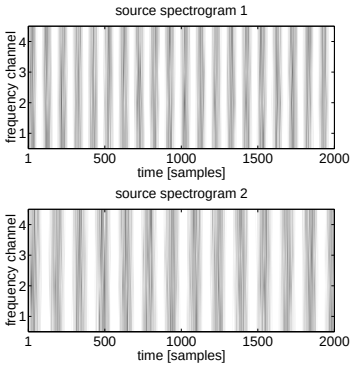

Figure 3: Spectrograms generated in experiment I.

solve this permutation ambiguity [1, 3, 4]. They consist of a two-step architecture: The source separation problem (Eq. (1)) is solved separately in each frequency channel $f = 1, \ldots, \nu$, and in a second step the reconstructed signals are sorted such that the ordering with respect to the original signals is the same in each frequency channel.

NOVEL ALGORITHM FOR BLIND SEPARATION OF MODULATED SIGNALS

In this section we present a new algorithm for the blind separation of convolutively mixed signals which is based on correlated modulation in different frequency channels of the source signals. For example, speech contains correlated modulation in different frequency regions. In contrast to existing algorithms, our algorithm does not use higher-order statistics or time-delayed correlations. The algorithm reconstructs source signals with the same ordering in each frequency channel. Hence, there is no need for a two-stage architecture.

Speech can be regarded as a stochastic signal, which is the approach usually adopted in blind source separation algorithms. But speech does also exhibit a rich deterministic structure, which is particularly obvious from its spectrogram representation as shown in Fig. 1. We are interested in the strong modulations present in speech signals. These modulations in the different frequency channels of a speech signal are highly correlated, i.e., a change in level in one frequency channel coincides with a change in level in many other frequency channels, which can also be observed in the spectrogram.

The origin of this lies in the production of speech: A broadband sound is emitted from the glottis and filtered by the vocal tract. Any change in glottal excitation or in the geometry of the vocal tract alters the spectrum of the resulting speech signal not selectively at a certain frequency but across a wide spectral range.

We define the correlation $C\left(\hat{s}_{i,\alpha}, \hat{s}_{j,\beta}\right)$ between modulations in different frequency channels α and β of two (generally different) signals $s_i\left(t\right)$ and $s_j\left(t\right)$ as

$$C(\hat{s}_{i,\alpha}, \hat{s}_{j,\beta}) \equiv \left\langle \left(|\hat{s}_{i,\alpha}(T)| - \langle|\hat{s}_{i,\alpha}(T)|\rangle_T\right)\left(|\hat{s}_{j,\beta}(T)| - \langle|\hat{s}_{j,\beta}(T)|\rangle_T\right) \right\rangle_T \tag{2}$$

By $\langle\cdot\rangle_T$ we denote the expectation value with respect to time and $|\cdot|$ denotes the norm of a complex number. Computed from a single source, $i = j$, the correlation $C\left(\hat{s}_{i,\alpha}, \hat{s}_{i,\beta}\right)$ is non-zero for almost all frequency pairs (α, β) (refer to Fig. 2). For two independent sources, $i \neq j$, it is clear that $C\left(\hat{s}_{i,\alpha}, \hat{s}_{j,\beta}\right) = 0$ for all pairs (α, β).

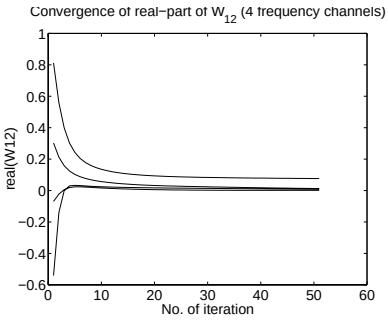

Figure 4: Convergence of parameter real$([\hat{W}_\alpha]_{12})$ with $\alpha = 1 \ldots 4$; permutations do not occur.

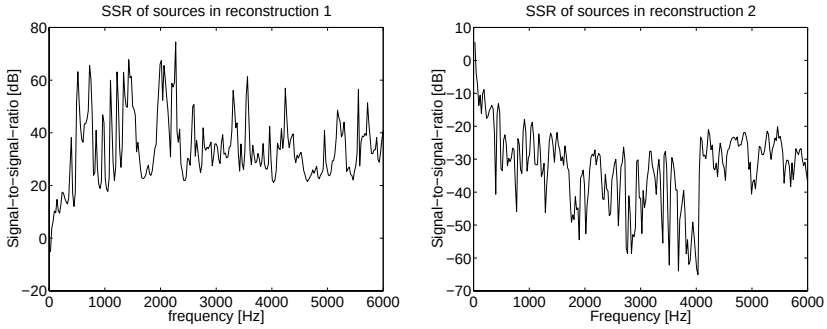

Figure 5: Signal-to-signal-ratio, i.e., ratio of source signal energies (first vs. second source) present in the first (left) and the second (right) reconstructed signal, respectively.

To obtain an optimisation algorithm, we define the cost function $E(\{\hat{\mathbf{W}}_\alpha\})$:

$$E\left(\{\hat{\mathbf{W}}_\alpha\}\right) = \sum_{i,j\neq i,\alpha,\beta} \left[C\left(\hat{x}_{i,\alpha}, \hat{x}_{j,\beta}\right)\right]^2, \tag{3}$$

where the reconstructed signals are defined as $\hat{\mathbf{x}}_\alpha(T) = \left[\hat{x}_{1,\alpha}(T), \ldots, \hat{x}_{N,\alpha}(T)\right]^T = \hat{\mathbf{W}}_\alpha \hat{\mathbf{m}}_\alpha(T)$. The cost function can be minimised by gradient descent. The minimum is acquired if the modulation in each frequency channel α of each source i is decorrelated with each frequency channel β of every different source $j \neq i$. As shown in the next section, this separates the sources. Clearly, at the cost function's minimum the order of the sources' components in the reconstructed signals $\hat{\mathbf{x}}_\alpha(T)$ is the same for every frequency channel α; any permutation would introduce correlations which would increase $E(\{\hat{\mathbf{W}}_\alpha\})$.
In practice, optimisation is improved by whitening the data and minimising Eq. (3) for whitened data.

EXPERIMENT I: ARTIFICIAL DATA

In order to demonstrate that our algorithm uses different modulations to separate sources, we performed a separation experiment with artificially generated data.
Two spectrograms with 4 frequency channels and 2000 time points were generated in the following way. Real and imaginary part of their components $\hat{s}_{i,\alpha}(T)$ were chosen randomly from a Gaussian distribution with zero mean and unit variance. The signals were sinusoidally modulated in each frequency channel. The modulation frequency was $(100 \text{ samples})^{-1}$ for the first signal and $(150 \text{ samples})^{-1}$ for the second signal, respectively. Modulation depth was 1 for both signals (refer to Fig. 3).
The signals were mixed with the same mixing matrix for each frequency channel, $\hat{\mathbf{A}} = \begin{pmatrix} 1 & 1 \\ \sqrt{-1} & 1 \end{pmatrix}$.
The elements of the separating matrices $\hat{\mathbf{W}}_\alpha$ were initialised differently for each frequency α, with real and imaginary part chosen randomly between -1 and 1. Due to the possible permutations, the algorithm should converge at one of the two possible solutions which, after scaling the diagonal elements to unity, are $\hat{\mathbf{W}}_\alpha = \begin{pmatrix} 1 & -1 \\ -\sqrt{-1} & 1 \end{pmatrix}$ and $\hat{\mathbf{W}}_\alpha = \begin{pmatrix} 1 & \sqrt{-1} \\ -1 & 1 \end{pmatrix}$, for all α.

After 50 iterations of the algorithm, the source signals were reconstructed successfully with an overall signal-to-signal-ratio of more than 25dB in the reconstructed signals. The algorithm converged at the same of the two possible solutions in each frequency channel, i.e., no permutations occurred (refer to Fig. 4). As expected, the algorithm failed to separate unmodulated signals.

EXPERIMENT II: SPEECH SIGNALS

By separating digital mixes and real-world recordings of speech signals, we demonstrate that the proposed algorithm can be applied to speech.
Spectrograms of two speech signals (sampling rate: 12kHz) of 4.5s duration were computed using a 256 samples long Hanning window, a shift of 64 samples and a DFT length of 512 samples. They were digitally mixed using a different mixing matrix $\hat{\mathbf{A}}_\alpha$ for each frequency α, composed of randomly chosen elements with real and imaginary part between -1 and 1. The separating matrices $\hat{\mathbf{W}}_\alpha$ were initialised with the unit-matrix for each frequency α. Expectation values in Eq. (2) were computed as the signals' means.
The algorithm reconstructed the source signals successfully. The overall signal-to-signal-ratio was 17dB in the reconstructed signals. From the frequency dependent signal-to-signal-ratio (refer to Fig. 5) it is clear that the order of the sources was the same in each frequency channel of the reconstructed signals.
In first experiments with noisy real-world stereo recordings of speech the algorithm achieved a signal-to-signal-ratio of more than 13dB.

CONCLUSION

A new algorithm is proposed for blind source separation of signals which contain correlated modulation in different frequency channels. As shown in the first experiment, the algorithm successfully separates artificially generated signals which contain correlated modulation. Hence, correlated modulation is a criterion which can be exploited by blind source separation algorithms. In the second experiment we demonstrate that the algorithm can be applied to speech signals. Speech contains correlated modulation in different frequency regions. The signal-to-signal-ratio in the separated signals is 17dB for digitally mixed speech and more than 13dB for real-world stereo recordings.

ACKNOWLEDGEMENTS

The author would like to thank Torsten Dau for suggesting the use of modulated artificial signals.

REFERENCES

[1] Murata, N., Ikeda, S., Ziehe, A., Tech. Rep. 98-2, Riken Brain Science Institute, Tokyo, 1998.

[2] Nadal, J.-P., and Parga, N., *Neural Computation* **9**, pp. 1421–1456 (1997).

[3] Parra, L., Spence, C., De Vries, B., Convolutive blind source separation based on multiple decorrelation, in *IEEE Neural Networks for Signal Processing Workshop*, Cambridge, UK, 1998.

[4] Servière, Ch., Blind source separation in presence of spatially correlated noises, in *ICA '99 Workshop*, Aussois, France, pp. 497–502, 1999.