# BLIND SEPARATION OF SPEECH SIGNALS

K. J. Pope[†] and R. E. Bogner

[†]Cooperative Research Centre for Sensor Signal
and Information Processing

Department of Electrical and Electronic Engineering,
University of Adelaide

ABSTRACT — Blind signal separation deals with the problem of extracting independent source signals, such as different speech signals, from a set of measurements that are combinations of the source signals, such as measurements made by several microphones in a room with several speakers. Applications that involve this problem are hearing aids, noise cancellation in extreme environments (e.g. aircraft cockpits and factory floors), and eliminating cross-talk between communication channels.

## INTRODUCTION

Recent research has shown that combinations of signals can, in principle, be processed to recover the original signals, under the assumption that the original signals are independent. There are two basic ways of distinguishing different source signals: temporal correlations and higher-order statistics (Pope and Bogner, 1994). Denoting the source signals by $s_i(n)$ and the measured signals by $x_i(n)$, the aim is to produce a set of output signals $y_i(n)$ that are independent. This implies that generalised cross-correlations of the output signals are separable, i.e.

$$E[f(y_i(n))g(y_j(n-k))] = E[f(y_i(n))]E[g(y_j(n-k))] \tag{1}$$

where $f(.)$ and $g(.)$ are arbitrary non-linear functions and $k$ is an integer time lag. The non-linear functions introduce the higher-order statistics, and the time lag $k$ introduces the temporal correlations. Equation (1) can be used as a test for independence, or it can be rewritten in the equivalent form of generalised covariances (Bogner, 1992):

$$E[(f(y_i(n)) - E[f(y_i(n))])\,(g(y_j(n-k)) - E[g(y_j(n-k))])] = 0 \tag{2}$$

Sources with identical spectra cannot be distinguished by temporal correlations, and hence cannot be separated by a method that only uses temporal correlations. Similarly, higher-order statistics cannot distinguish sources that generate Gaussian-distributed data. Hence methods that use only higher-order statistics cannot separate Gaussian sources. Using both types of information, only Gaussian sources with identical spectra cannot be separated. Various algorithms have been proposed in the literature, mostly based on higher-order statistics. (Weinstein et al., 1993) notes the possibility of using both temporal correlations and higher-order statistics in a separation algorithm. (Yellin and Weinstein, 1993) describes a frequency-domain algorithm for two input and two outputs based on third-order statistics. Whilst this algorithm can only separate sources which are distributed asymmetrically, it can in principle be adapted to more general problems.

After stating the problem mathematically, we propose conceptually simple solution to the problem is proposed, based on the minimisation of a cost function. The cost function combines many tests for independence using generalised covariances, and its minimisation yields the parameters of the separation system. Simulations with speech data illustrating the use of the algorithm are given, and conclusions are drawn with suggestions for future work.

## STATEMENT OF THE PROBLEM

Let $s_i(k)$ be the $i$th source signal at time $k$, and the concatenation of the $n$ source signals be the vector $s(k) = [s_1(k), \dots s_n(k)]^{\mathrm{T}}$. Similarly, the measurements made at the $m$ sensors are denoted $x_i(k)$, and are collected to form the vector $x(k) = [x_1(k), \dots x_m(k)]^{\mathrm{T}}$. The standard blind signal separation problem considers linear, instantaneous, time-invariant combinations of source signals, and hence $s(k)$ and $x(k)$ are related by a constant matrix multiplication

$$x(k) = As(k)$$

where $A$ is an $m \times n$ matrix.

Separation is effected by a similar matrix multiplication. Let the outputs of the separation system be $y_i(k)$, concatenated into the vector $y(k)$, and the $n \times m$ separation matrix be $B$. Hence

$$y(k) = Bx(k)$$

It can be shown (Comon, 1994) that separation is achieved if $AB = DP$ where $D$ is a diagonal matrix, and $P$ is a permutation matrix. The overall separation system is shown in figure 1.

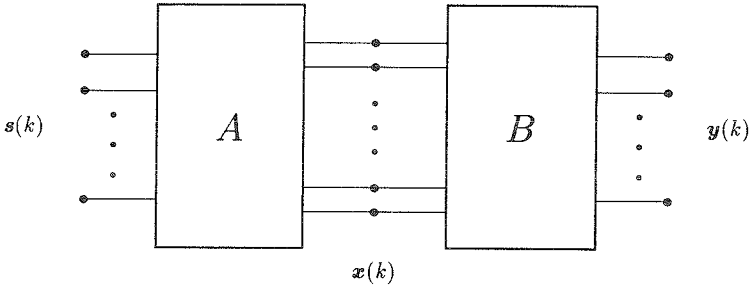

Figure 1: Signal separation system

## SOLUTION METHOD

A common method for solving a problem of this nature is to find the extremum of some cost function. In this case, the cost function should measure how far the outputs are from being independent. As detailed in equations (2) and (1) in the introduction, a generalised covariance or cross-correlation can be used to test whether the system outputs are independent. One simplification that reduces the computational load is to use a subset of generalised cross-correlations of the form

$$E[y_i(n))g(y_j(n-k))]$$

If the outputs have a mean of zero, then the above generalised cross-correlations will equal zero when the outputs are independent.

A cost function can be built by summing the squares of several of these generalised cross-correlations over all pairs of inputs and several time lags, i.e.

$$Q = \sum_{\substack{i,j=1 \\ j \neq i}}^{P} \sum_{k=0}^{L} \sum_{l=1}^{G} E[y_i(n) g_l(y_j(n-k))]^2 \tag{3}$$

There can easily be many terms in the summation, and choosing which terms to use will often be dictated by implementation issues. This formulation is particularly amenable to a parallel implementation, as each generalised cross-correlation can be calculated independently of the others.

### System structure

The implementation of any signal separation algorithm will be of one of two types. The first is a block-based solution, where the separation system's parameters are held constant over some block of time, and are determined principally from the measurements made in that time block. The second method is an iterative solution, where the parameters are continually updated as new measurements become available. The two methods can be compared by considering the memory of the two systems. In the block-based approach, time is divided into blocks. The system has perfect memory of samples within the current block, but no memory of samples outside the block. In the recursive approach, the system has perfect knowledge of the present, but its memory decays (typically exponentially) into the past.

The block-based implementation typically uses a stationarity assumption within each block, and hence the separation quality will be degraded if the inputs are non-stationary. In addition, the outputs of the system may exhibit 'click' artefacts at the block boundaries due to abrupt changes in parameter values. This may be alleviated by overlapping blocks, but at the expense of extra computation. In the extreme case, the changes in parameter values may result in cross-over, where an output stops tracking a particular source, and jumps to another source. In contrast, a recursive implementation will exhibit no repetitive artefacts, but may suffer from slow convergence to separation or poor tracking of non-stationary features (for example, caused by a moving source or sensor). There will also be some additional parameters, such as the learning rate, that must be determined. It may be desirable to vary these parameters, in which case a method for varying their values must also be chosen.

### Implementation details

The cost function minimisation can be implemented in a variety of ways. A highly efficient implementation for a block-based algorithm is the Joint Approximate Diagonalisation of Eigenmatrices (JADE) algorithm (Cardoso and Souloumiac, 1993). Although the original algorithm utilises only higher-order statistics, a version that uses only temporal correlations was presented in (Abed-Meraim et al., 1994). Implementation details of cost function minimisation algorithms for both a block-based system and an adaptive system using a neural network to separate the sources are given in (Burel, 1992). For linear separation systems, a simpler algorithm is obtained.

## SIMULATIONS

The results given in this section use a synthesis of the JADE algorithms described in (Cardoso and Souloumiac, 1993; Abed-Meraim et al., 1994). The algorithm is block-based, and a single block is used. The source signals used are 2.5 seconds of speech or Gaussian noise, sampled at 8kHz. All the simulations reported here use synthetic combinations of source signals.

### Noise cancellation

To investigate the separation system's ability to reject a noise source and separate the desired speech signals, a system with three sources, three sensors and three outputs was simulated. This is an extension of the classic noise cancellation problem, but with significant signal leakage into the noise reference signal. Two of the sources were speech signals of approximately equal energy, and the third was white Gaussian noise with an energy varying from -40dB to +40dB (relative to the speech sources) over several experiments. The combination matrix $A$ was set to be

$$A = \begin{bmatrix} 1.00 & 0.99 & 0.95 \\ 0.98 & 1.00 & 0.93 \\ 0.92 & 0.99 & 1.00 \end{bmatrix}$$

This results in sensor signals $x_1(k)$, $x_2(k)$ and $x_3(k)$ that are indistinguishable by ear. The matrix is poorly conditioned, making the separation task very difficult.

Figure 2 plots the input and output Signal-to-Noise Ratios (SNR) for both speech signals for a range of noise source energies. The input SNR is defined by comparing the energy of that component of the sensor signal that is identical to the desired source signal with the energy of the remainder of the sensor signal. A similar definition holds for the output SNR, based on the output signal rather than the sensor signal. The noise source energy is measured in dB relative to the source signal energy; hence 0dB noise source energy specifies that the noise source has the same energy as the signal. Notice that the output SNR is independent of the energy of the interfering noise source, as the noise source can be exactly cancelled. The output SNR, in this example, is in fact limited only by numerical precision.

### Ill-conditioned sensor signals

When the sensor signals are nearly identical, the separation system becomes extremely sensitive to measurement noise. To illustrate this, two speech sources of approximately equal energy were combined by the matrix

$$A = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix}$$

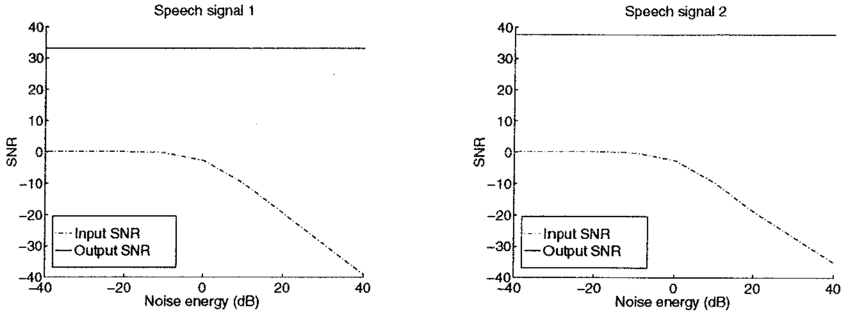

Figure 2: Input and Output SNR against input noise energy

Independent noise of energy from -40dB to +10dB relative to the speech signals was added to each sensor signal.

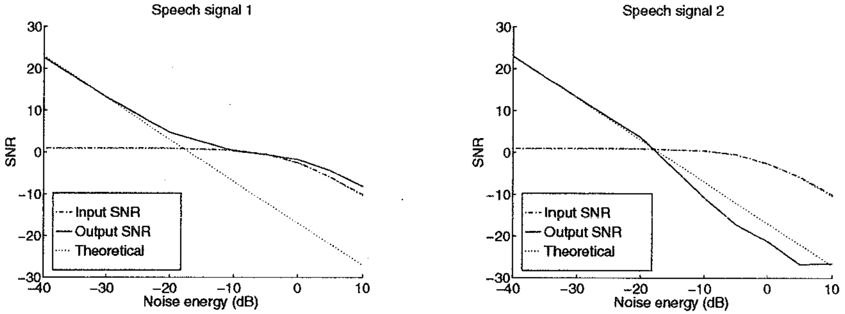

Figure 3: Input and Output SNR against input noise energy

Figure 3 shows the input and output SNRs for each speech signal, along with the theoretical values for these quantities. It is clear that the maximum SNR attainable is significantly less than the simple ratio of the signal energy to the noise energy. This is because the cancellation of the interfering speech signal requires taking the difference between two large quantities. Thus the amplitude of the desired speech signal is small, giving a small energy. As the noise in the two sensors is uncorrelated, the difference of the two signals results, on average, in a noise signal with energy equal to the sum of the individual energies. Hence for the system at hand, the noise is boosted by 17dB.

It is also clear that as the noise energy rises beyond -20dB, the system fails to separate the sources as predicted by the theory. The first output (the left-hand plot) is an approximately equal weighting of the two sensor signals. As both speech signals reinforce coherently whereas the noise signals simply combine their energies, the output SNR is approximately 3dB higher than the input SNR. In contrast, the second output does separate the second speech signal from the first, but at a significantly smaller amplitude than desired. The output noise energy is also reduced, but to a lesser extent. Hence the output SNR is less than theoretically predicted SNR.

It is possible that the signal separation fails because as the noise energy increases, the cost function that is being minimised becomes dominated by the statistics of the noise. Thus the global minimum of the cost function corresponds to signal separation at low noise energies, but as the noise energy increases, this minimum becomes shallower and shallower. It seems that as the input SNR and output SNR curves cross, this minimum becomes unstable.

## CONCLUSIONS

This paper has introduced the problem of blind signal separation, in the context of speech signals. A brief discussion of the theory and methods of solution have been given. Simulations have been presented to illustrate some of the ideas outlined, and to highlight how the separation algorithms work in practice.

Work in this area is quite new, and there are many limitations with the results presented here. Further investigations into blind signal separation should address the following problems.

- Many algorithms do not consider the effects of sensor noise, both independent noise at each sensor, and dependent noise measured by all sensors.
- In general, the number of sources present at any time will be unknown. This is especially true for noise sources.
- The assumption of stationarity or time-invariance used in several algorithms will not be accurate in many real-world situations.
- The assumption of linear instantaneous source combinations is similarly inappropriate in many situations. In particular, multi-path propagation or delay in the arrival of a signal at different sensors is expected.

### Future Work

In addition to the general problems listed above, some more specific areas of interest can be identified. A theoretical analysis of the local minima of the cost function should be undertaken, to identify the conditions under which an algorithm based on that cost function is guaranteed to converge to the desired solution. A closely related idea concerns the weighting of the generalised cross-correlations. In equation (3), several generalised cross-correlations are summed, which is a linear combination with unity weighting. It is worth investigating whether a different weighting results in a cost function with better characteristics, such as the number and size of local minimum, or sensitivity to noise characteristics.

Multi-path propagation is very important in many situations, but particularly so for blind separation of speech signals. To allow for this effect, the problem must be reformulated. Fortunately, transforming the extended problem to the frequency domain gives a form almost identical to the simple form considered in this paper. By applying simple separation algorithms independently at each frequency considered in the Fourier domain, signals that have propagated via several paths can in principle be separated.

## REFERENCES

Abed-Meraim, K., Belouchrani, A., Cardoso, J. F., and Moulines, E. (1994). Asymptotic performance of second order blind separation. In *International Conference of Acoustics, Speech and Signal Processing*, volume IV, pages 277–280.

Bogner, R. E. (1992). Blind separation of sources. Technical Report 4559, Defence Research Agency, Malvern.

Burel, G. (1992). Blind separation of sources - A nonlinear neural algorithm. *Neural Networks*, 5(6):937–947.

Cardoso, J. F. and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings, Part F*, 140(6):362–370.

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, 36:287–314.

Pope, K. J. and Bogner, R. E. (1994). Blind signal separation. submitted to *Digital Signal Processing*.

Weinstein, E., Feder, M., and Oppenheim, A. V. (1993). Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, 1(4):405–413.

Yellin, D. and Weinstein, E. (1993). Multi-channel signal separation based on cross-bispectra. In *IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 270–274.

# VOICE SEPARATION BASED ON MULTI-CHANNEL CORRELATION AND COMPONENTS TRACING

Jie Huang† and Noboru Ohnishi†‡

†Bio-Mimetic Control Research Center
The Institute of Physical and Chemical Research (RIKEN)

‡Department of Information Engineering
Faculty of Engineering, Nagoya University

ABSTRACT - This paper presents a novel sound separation method for voices pronounced by multiple persons from different directions. The method uses multi-channel signals from microphones located on different positions. Signals from all microphones are divided into narrow sub-bands by a set of band-pass filter. Envelope section curves over frequencies are calculated at each decimated sampling point. We assume that the voice energy meanly exists in the peaks of envelope section curves. The separation task, then, is to group all of those peaks. Arrival temporal disparities are used to group peaks in the method. Grouping operation, in order to cope with echo, is applied only when a peak just appears. It is because reflected sound arrives later than direct sound, and thus onset only contains sound directly from its source. The peaks are traced after onset and the grouping is maintained. Voice separation experiments were conducted in both an anechoic chamber and a normal room enclosed by concrete walls. The availability of the method was demonstrated.

## INTRODUCTION

The ability of a human to separate a particular sound from a noisy environment is called the "cocktail party effect" (Cherry 1953). Some proposed speech separation methods utilize monaural cues like the harmonic and synchronous characteristics of speech (Parsons 1976, Cooke 1992). A single monaural cue, however, is usually not perfectly sufficient for sound separation. In speech only vowels have harmonic characteristic, and synchronous will not be maintained in all frequency components. It is well know that the "cocktail party effect" is more effective when listening binaurally than listening monaurally, i.e., multiple channel signal is more effective than single channel signal. However, multi-channel based sound separation methods usually focus on sound parameters estimation by using total sound components. In a very complicated environment, they need a large computation power and are difficult to build a mathematical model. They will be not robust against noise, and are difficult to adapt to multi-path environments.

Our approach is to segment sound into small pieces in both time and frequency domain and then group them into several groups by the difference of arrival temporal disparities (ATD) between different microphones. Each sound piece contains narrow band frequency component and has an onset as starting point and an offset as ending point. A new method named as Group Onset Trace Ongoing Method (GOTO Method) was proposed. It groups sound pieces at onsets by ATDs and then traces the peaks of envelope section curve to keep the grouping. The benefit of sound segmentation is each sound piece can be considered as coming from single sound source; and grouping sound pieces by onset ATDs can eliminate the adverse influence of echoes.

## GROUP ONSET TRACE ONGOING METHOD

### Envelope Section Curve and Peak Detection

Sound signals are past to a set of narrow band-pass filter. Envelope of each sub-band signal