Technical Report: OSU-CISRC-6/99-TR15
Department of Computer and Information Science
The Ohio State University
Columbus, Ohio 43210-1277

A Comparison of Auditory and Blind Separation Techniques for Speech Segregation

André J. W. van der Kouwe, DeLiang L. Wang, Guy J. Brown

Abstract

A fundamental problem in auditory and speech processing is the segregation of speech from concurrent sounds. This problem has been a focus of study in computational auditory scene analysis (CASA), and it has also been recently investigated from the perspective of blind source separation. Using a standard corpus of voiced speech mixed with interfering sounds, we report a comparison between CASA and blind source separation techniques, which have been developed independently. Our comparison reveals that they perform well under very different conditions. A number of conclusions are drawn with respect to their relative strengths and weaknesses in speech segregation applications as well as in modeling auditory function.

A. J. W. van der Kouwe is with the Biomedical Engineering Center, The Ohio State University, Columbus, OH 43210. Email: avdk@ieee.org.

D. L. Wang is with the Department of Computer and Information Science and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210. Email: dwang@cis.ohio-state.edu.

G. J. Brown is with the Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Streeet, Sheffield S1 4DP, U.K.. Email: g.brown@dcs.shef.ac.uk.

I. Introduction

Human listeners exhibit a remarkable ability to segregate the voice of a single speaker from a mixture of other intruding sounds. This phenomenon may be regarded as one aspect of a more general process of auditory organisation, which is able to untangle an acoustic mixture in order to retrieve a perceptual description of each constituent sound source. The term auditory scene analysis (ASA) has been introduced to describe this process [3]. Conceptually, ASA may be regarded as having two stages. In the first stage, the acoustic mixture is decomposed into sensory elements ('segments'). The second stage ('grouping') then combines segments that are likely to have originated from the same sound source.

Recently, attempts to develop computational systems that mimic ASA have led to the emergence of a new field, known as computational auditory scene analysis (CASA) [4], [17]. Much of the work in this field has focused on the problem of segregating speech from interfering sounds ('speech segregation'). An effective computational solution to this problem would have application, for example, as the front-end to an automatic speech recognition system or as the basis for a hearing prosthesis.

The problem of speech segregation has also received attention from workers investigating blind source separation [13]. In contrast to CASA, blind source separation is a statistical technique that draws no inspiration from mechanisms of auditory function - nonetheless, it has been used to separate mixtures of audio signals with some success [10], [2].

The purpose of this paper is to compare the CASA approach to speech segregation with the blind source separation approach. We choose a representative technique from each domain and apply it to a standard corpus of speech mixed with various forms of interfering noise. Our study is motivated by a desire to determine the conditions under which it is most appropriate to apply each approach. Additionally, we aim to identify the limitations of the two techniques, and hence to suggest directions for future research.

To represent CASA we have chosen the system of Wang and Brown [19]. Blind source separation is represented by the Joint Approximate Diagonalization of Eigen-matrices (JADE) algorithm of Cardoso et al. [7]. The CASA and blind source separation domains are reviewed in Sections II and III respectively, and in these sections we also give reasons for our choice of representative techniques. Section IV describes the data set used for the evaluation. Finally, we discuss the relative performance, benefits and shortcomings of the two techniques in Section V.

II. Computational Auditory Scene Analysis

One of the first approaches to CASA is a system for separating simultaneous talkers described by Weintraub [20]. In his system, a frequency analysis of the acoustic mixture is performed by a bank of bandpass filters, and the interpeak interval in each filter channel is determined. This information is used to estimate the number of sound sources present, and their pitch periods. Each voice is characterized by the state of a Markov model - silent, periodic, non-periodic, onset, offset, increasing periodicity or decreasing periodicity. A spectral estimation algorithm then uses information about the state of each sound source to determine how the energy in each frequency channel should be allocated.

A later approach described by Brown and Cooke [4] addresses some of the problems of early CASA techniques; in particular, it avoids making strong assumptions about the type and number of sound sources. Additionally, their model attempts to put biological realism on a firmer footing. Their approach includes a model of the auditory periphery consisting of outer and middle ear filtering, cochlear filtering and a model of neuromechanical transduction. Central processing is modelled using various auditory maps, which represent periodicities, frequency transitions, onsets

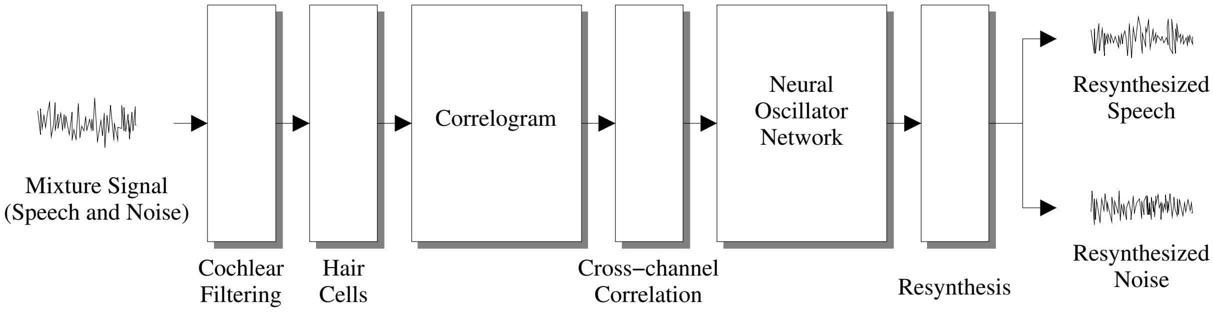

Fig. 1. Schematic diagram of the Wang and Brown CASA model. The input first passes through a model of the auditory periphery (cochlear filtering and hair cells) which simulates auditory nerve activity. Mid-level auditory representations are then formed (correlogram and cross-channel correlation map). Next, a two-layer neural oscillator network performs grouping of acoustic components. A final resynthesis path facilitates computation of signal-to-noise ratio (adapted from [19]).

and offsets. These maps correspond with observed neural mappings in the auditory mid-brain and cortex. The auditory scene is divided into symbolic elements using the maps, and elements are grouped according to the similarity of their various features.

Wang and Brown [19] extend the work of Brown and Cooke by replacing the central processing portion of their model with a two-layer oscillator network and employing computationally simpler methods for auditory feature extraction. Segmentation of the acoustic input arises from the dynamics of the first layer, through a process of local excitation and global inhibition, while grouping of segments emerges from the dynamics of the second layer. A population of synchronized oscillators represents an individual sound source. Different sources are represented by desynchronized populations of oscillators. This 'oscillatory correlation' framework represents a solution to the binding problem of how distributed sensory components of a single source are bound together in the brain, and is supported by recent neurobiological findings. The system is illustrated schematically in Figure 1. Note that the system allows a time-domain waveform to be resynthesized for each segregated sound source.

The Wang and Brown system has been chosen to represent CASA techniques for four reasons. Firstly, it is strongly motivated by auditory neurobiology; in contrast, many other CASA approaches are inspired by mechanisms of auditory function but do not model them closely (e.g., [11], [15]). Secondly, it performs well and the results have been recently published. Thirdly, the performance of the system has been tested using a readily-available corpus of sound mixtures, which is described in Section IV. Finally, since a waveform can be resynthesized for the separated speech and noise, it is possible to express the performance of the system using a commonly used metric - signal to noise ratio (SNR). This allows a direct comparison with the performance of the JADE algorithm.

III. Blind Source Separation

Blind source separation relies on the availability of several differing source mixtures and relatively strict requirements on the statistical properties of the sources. When these conditions are satisfied the technique is capable of near perfect source separation. By definition, in blind separation there is no available a priori knowledge as to the statistical distributions of the source signals. Nor is there information available as to the nature of the process by which the source signals were combined. Therefore some assumptions must be made regarding the source signal distributions and a model of the mixing process must be adopted. It is generally assumed that the source signals are statistically independent and that the mixing process is linear. Blind separation algorithms attempt to invert the mixing process in such a way as to recover components which are in some sense independent.

Statistical independence implies that the source signal joint moments of all orders are zero. Algorithms ensuring only that the second order joint moments are zero (i.e., that the covariance matrix is unit) fall into the category of principal component analysis (PCA). Similarly, algorithms that operate explicitly on higher-than second order statistics are classified as independent component analysis (ICA) (we note that some nonlinear PCA algorithms implicity operate on higher order statistics [13]). For example, the ICA procedure of Comon [8] minimizes the fourth-order cumulants given by (1) after whitening [1] the signals.

$$c_{ICA}[\overline{y}] = \sum_{ijkl \neq iiii} |\widehat{cum}(y_i, y_j, y_k, y_l)|^2 \tag{1}$$

where $\overline{y}$ is the vector of approximately separated signals. The Joint Approximate Diagonalization of Eigen-matrices (JADE) algorithm [6] minimizes the cumulant given by:

$$c_{JADE}[\overline{y}] = \sum_{ijkl \neq ijkk} |\widehat{cum}(y_i, y_j, y_k, y_l)|^2 \tag{2}$$

It has been shown that the performance of algorithms that implement (1) and (2) are equivalent, but a faster optimization process exists for JADE [7].

The simplest model for the mixing process is [14]:

$$\overline{m}(t) = \mathbf{M}\overline{s}(t) \tag{3}$$

where $\overline{s}$ is the vector of original source signals, $\mathbf{M}$ is the mixing matrix and $\overline{m}$ is the vector of mixed signals observed by the microphones. This simple linear model aligns the source signals perfectly in time in the mixtures. For real audio signals this is very unlikely due to the different path lengths from the sources to the various microphones [18]. Another complication in real acoustic environments is signal distortion by filtering and echoes. Convolutive mixing models have been proposed to account for these effects and these may include delays implicitly or explicitly [10]:

$$m_i(t) = \sum_j (h_{ij} * s_j)(t)$$

where h is the filter impulse response. Finally, the mixing model may be nonlinear [21].

The existing blind source separation algorithms therefore vary in terms of three properties - the complexity of the mixing process model, the order of the statistics used by the algorithm and whether the algorithm processes the data iteratively or in batches. We have chosen the JADE algorithm based on these three properties. JADE assumes a linear mixing model. Although other algorithms incorporate delays and nonlinearities in the mixing model, we know that our test data set is constructed synthetically by linear mixing so that the linear model given by (3) is appropriate. JADE is a fourth order technique. Several algorithms are iterative, converging on a set of parameters for the separating model. JADE is not iterative, but acts on the statistics of the complete set of data directly. For separating linear mixtures of signals, JADE therefore represents the best that any equivalent fourth order iterative method can do.

[1] 'Whitening' refers to the process which transforms a signal vector so that the covariance matrix is unit.

IV. Test Data

We use the corpus of sound mixtures devised by Cooke [9] for our evaluations. It consists of 100 speech and noise mixtures, formed by combining each of 10 voiced utterances ($v0$ to $v9$) with each of 10 intrusions ($n0$ to $n9$) in the ratio 1:1. The intrusions, characterized in Table I, reflect the various types of intruding noise which may occur in a natural listening environment. Five sentences were spoken by two male speakers to obtain the ten fully voiced utterances. These are listed in Table II.

TABLE I
Intrusive sources (from Cooke, 1993).

ID	Description	Characteristics
$n0$	1 kHz tone	narrowband, continuous, structured
$n1$	white noise	wideband, continuous, unstructured
$n2$	series of brief noise bursts	wideband, interrupted, unstructured
$n3$	teaching laboratory noise	wideband, continuous, partly structured
$n4$	new wave music	wideband, continuous, structured
$n5$	FM signal ("siren")	locally narrowband, continuous, structured
$n6$	telephone	wideband, interrupted, structured
$n7$	female TIMIT utterance	wideband, continuous, structured
$n8$	male TIMIT utterance	wideband, continuous, structured
$n9$	female utterance	wideband, continuous, structured

For blind separation, one signal and noise mixture per separation is insufficient, since the required number of mixture signals must equal or exceed the number of source signals. It has been pointed out that with carefully placed microphones, any single desired source signal may be obtained with fewer microphones than sources, at least in principle [5]. In practice this would require unrealistically precise placement of the microphones.

TABLE II
Voiced sources (from Cooke, 1993).

ID	Speaker	Utterance
$v0$	1	I'll willingly marry Marilyn
$v1$	1	Why were you away a year, Roy?
$v2$	1	Why were you weary?
$v3$	1	Why were you all weary?
$v4$	1	Our lawyer will allow your rule
$v5$	2	I'll willingly marry Marilyn
$v6$	2	Why were you away a year, Roy?
$v7$	2	Why were you weary?
$v8$	2	Why were you all weary?
$v9$	2	Our lawyer will allow your rule

In this experiment, 100 mixture *pairs* were created by combining the voice and intrusion signals in the ratio 1.0:0.8 and 0.8:1.0. It is important to note that the blind separation algorithm not only has two signals to work with in each case, but also has the implicit a priori knowledge that

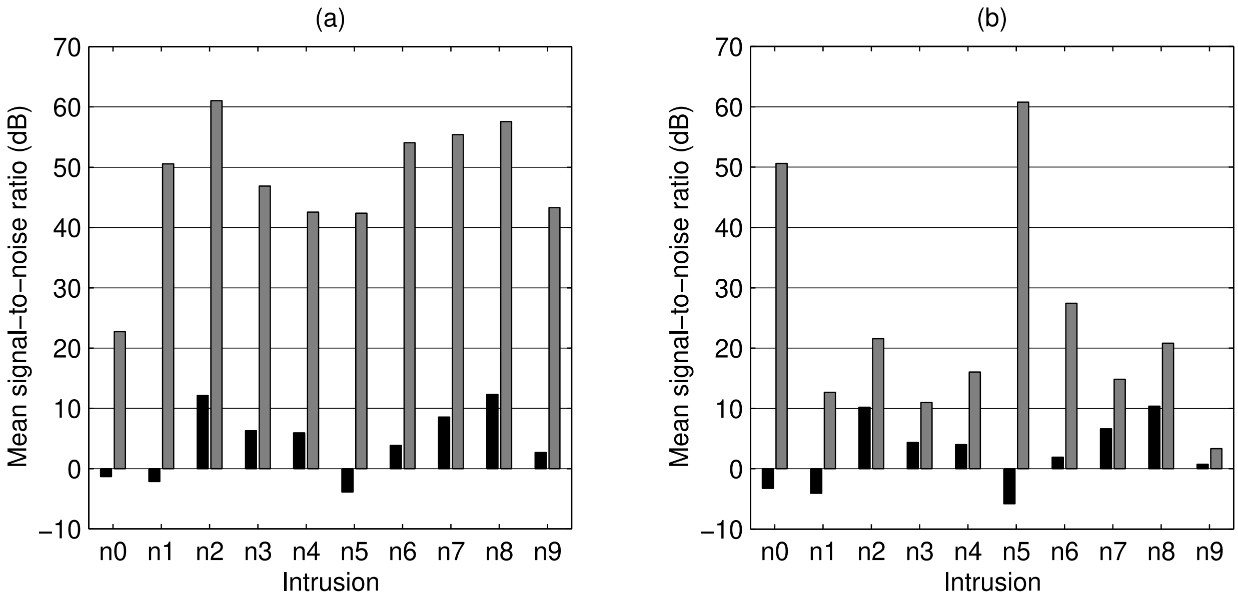

Fig. 2. Comparison of signal-to-noise ratio before and after segregation by (a) blind separation (JADE) and (b) computational auditory scene analysis. Voiced speech is segregated from a mixture of speech and ten different intrusions (n0 = 1 kHz tone; n1 = random noise; n2 = noise bursts; n3 = 'cocktail party' noise; n4 = rock music; n5 = siren; n6 = trill telephone; n7 = female speech; n8 = male speech; n9 = female speech) (from [19]).

the mixture consists of exactly two components. The results of this experiment with JADE are contrasted with those of the CASA algorithm of Wang and Brown [19]. For each type of intruding noise, the signal to noise ratio before and after separation is given. These signal to noise ratios are averaged over the 10 voice signals after conversion to decibels.

In contrast with the two mixtures that JADE had to work with, Wang and Brown presented their system with only a single mixture of signals, of which the signal-to-noise ratio (SNR) was equal to the average of the two presented to JADE. The SNR values appearing in Figure 2 for the mixtures before separation are therefore slightly higher for the blind separation technique than for the CASA technique, because the higher SNR of the two mixtures available in each case (the 1.0:0.8 mixture) were chosen for a fair comparison.

The SNR improvement in the case of the JADE algorithm can be calculated directly from the estimate of the unmixing matrix $\mathbf{N}$ and the true mixing matrix $\mathbf{M}$ if the unmixing model

$$\overline{y}(t) = \mathbf{N}^{-1}\overline{m}(t) \tag{4}$$

which is the reverse of (3) is adopted. The SNR is then

$$SNR = \left[\frac{n_{11}m_{00} - n_{01}m_{10}}{n_{11}m_{01} - n_{01}m_{11}}\right]^2 \tag{5}$$

The chosen CASA technique provides a resynthesis path for the separated signals [19], making it possible to determine the SNR improvement. The performance of the blind separation and CASA approaches may thus be compared directly in terms of signal-to-noise ratios.

V. Comparisons

The JADE results for each voice-intrusion pair are listed in Table IV along with the average over all voices for each intrusion type. The averages are calculated after conversion to decibels. The average values are shown in Figure 2(a). For comparison, the Wang and Brown results [19] are shown in Figure 2(b).

In order to understand these results and to be able to compare them meaningfully, it is necessary to contrast the requirements of JADE and the CASA technique.

- For JADE to be applied successfully, there are several requirements on the properties of the signals. These are summarized in Table III. Various blind separation algorithms other than JADE allow compromises on some of these requirements. All of them require that the source signals be in some sense statistically independent.
- CASA techniques in general require that the mixture signal, represented in the time-frequency plane, show well-defined regions corresponding to one or more of the source signals.

TABLE III

Source signal properties relevant to blind separation.

No.	Property
1	Statistical independence
2	Stationarity of sources
3	Constancy of mixing process
4	Linear mixing
5	Mixing matrix not near singular
6	Known, fixed number of sources
7	Equal number of sources and available mixtures
8	Perfect temporal alignment
9	No convolutive processes before mixing

The CASA approach imposes less stringent requirements on the statistical and mixing properties of the signals than the blind separation algorithms. The most important of these are properties 6 and 7 in Table III - blind separation requires that the number of sources be known and that the number of available mixture signals be equal to the number of sources, whereas CASA techniques require only a single mixture signal. In our experiment with JADE, we have satisfied all of these requirements by design, except for the requirement of statistical independence of the source signals. We have a constant, linear mixing process, a mixing matrix which is far from singular, two sources and two mixture signals, and perfect temporal alignment of the source signals within the undistorted mixtures.

For separating sounds in a natural auditory environment, it is important to consider whether these requirements on the statistical and mixing properties of the signals are realistic.

- Echoes give rise to distortions in the representation of the source signals in the mixtures. Differing path lengths to the microphones result in temporally misaligned signals [18]. Some blind separation algorithms handle distortions and delays by modeling convolutive processes [2]. Signals containing echoes have not been dealt with by CASA algorithms, but one would expect a lesser problem because segregation in CASA algorithms are based on intrinsic properties of auditory signals, such as pitch, and these properties tend to be preserved in echoes. Additionally, computational models of the precedence effect [12] could be incorporated into CASA systems in order to minimize the distortion caused by echoes.

TABLE IV
SNR IMPROVEMENT AFTER APPLYING JADE TO STANDARD DATA SET.

	$n0$	$n1$	$n2$	$n3$	$n4$	$n5$	$n6$	$n7$	$n8$	$n9$
$v0$	3.59	41.7	40.2	51.4	38.1	47.0	46.0	26.0	60.6	61.7
$v1$	48.9	47.0	40.7	35.2	57.7	49.9	63.8	69.6	50.3	22.9
$v2$	65.3	48.2	39.2	36.3	31.9	66.1	52.4	45.1	26.1	35.1
$v3$	39.8	46.1	32.0	25.3	27.6	44.7	45.9	18.6	28.6	46.9
$v4$	3.00	81.0	50.6	36.2	27.8	48.9	34.9	25.9	36.7	26.2
$v5$	4.41	50.6	31.9	31.7	36.6	38.5	46.2	37.0	25.4	60.4
$v6$	7.15	46.8	28.8	27.0	20.1	52.6	43.8	33.4	23.6	30.4
$v7$	54.5	58.9	31.2	18.1	16.7	51.5	43.5	56.2	29.4	29.1
$v8$	26.1	56.8	37.6	41.9	26.3	50.2	47.4	40.9	19.5	38.7
$v9$	1.02	71.0	36.1	39.9	24.2	52.0	40.0	30.2	29.5	28.2
Ave	25.4	54.8	36.8	34.3	30.7	50.1	46.4	38.3	33.0	38.0

- If the microphones are relatively close together, the effective mixing matrix (assuming that a linear mixing model is indeed suitable) may be near singular.
- If the sources move in space, the value of the corresponding mixing matrix varies with time. Batch mode algorithms such as JADE require that the source signals be stationary and that the mixing process not vary with time. The CASA algorithm has not been tested with moving sources, but, again, they are expected to be a lesser problem due to grouping cues that are intrinsic to the content of the source signals. Some blind separation algorithms, in particular the neural network approaches [14], *require* that the signals be non-stationary and may be able to handle a mixing process that varies slowly with time.

The results of our tests, shown in Figure 2, reveal the following:

- JADE performs best for the white noise ($n1$) and the siren ($n5$), and it performs worst for the 1 kHz tone ($n0$).
- The CASA technique performs better than blind separation only for the 1 kHz tone ($n0$) and the siren ($n5$), and performs worst for the female utterance ($n9$).

In principle, the blind separation problem is impossible to solve if more than one of the signals is Gaussian and mixing is linear, since the sum of Gaussian distributions is still Gaussian. The most fundamental assumption which must be made for blind signal separation to work is that the source signals are independent in some statistical sense. In JADE it is the statistics up to the fourth order which are used, and these statistics are required to be stationary. Since the other assumptions were satisfied by design of the experiment, it is the statistical properties alone which lead to these performance differences. This may be explained by the fact that the white noise mixtures contain rich higher-order joint statistics, whereas the 1 kHz tone mixtures yield poor higher-order joint statistics.

The greatest contributor to the variation in performance of the CASA technique over the range of intrusions is the structure of the intrusive sound in the time-frequency plane. This is characterised in terms of bandwidth, continuity and granular structure in Table I. Intrusions which are represented by a compact area in the time-frequency plan can be very effectively rejected by the Wang and Brown system, such as the 1 kHz tone and siren. Similarly, acoustic mixtures which are fragmented and overlapping in the time-frequency plane present the greatest challenge to the CASA technique. As a result, the CASA system performs comparatively poorly for broadband intrusions such as the random noise and speech. The particularly poor performance on $n9$ (the female utterance) may also reflect errors in the pitch tracking procedure.

Finally, it may be noted that humans are also limited in discerning many simultaneous auditory events. In such scenarios our strength lies in our ability to selectively attend to one event. This is well demonstrated in the cocktail party effect, where humans show a remarkable capacity for following one speaker in a noisy auditory scene [3]. In this sense, the performance profile of CASA shown in Figure 2(b) resembles human performance, whereas that of blind separation, shown in Figure 2(a), does not.

VI. Conclusion

Blind separation can be remarkably successful at separating mixtures of sounds, provided that certain strict requirements on the properties of the source signals are met. These requirements are often not equitable with a natural listening environment. In particular, the requirements that as many mixture signals as sources be available, and that they be perfectly aligned in time, are quite restrictive. If these conditions are met, blind separation is a powerful technique.

CASA algorithms are subject to quite different and biologically reasonable assumptions, and the performance profile for the test data set is correspondingly quite different. In the natural environment, the methods of computational auditory scene analysis bring the flexibility of the physiological systems which they model to bear on a variety of signal mixtures so that they can achieve a reasonable level of separation in the absence of many of the requirements of blind separation.

References

[1] S. Amari, T. Chen and A. Cichoki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks,* vol. 10, no. 8, pp. 1345-1351, 1997.
[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation,* vol. 7, pp. 1129-1159, 1995.
[3] A. S. Bregman, *Auditory Scene Analysis,* MIT Press, Cambridge, MA, 1990.
[4] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language,* vol. 8, no. 4, pp. 297-336, 1994.
[5] X. Cao, R. Liu, "General approach to blind source separation," *IEEE Transactions on Signal Processing,* vol. 44, no. 3, Mar. 1996.
[6] J. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F,* vol. 140, no. 6, pp. 362-370, Dec. 1993.
[7] J. F. Cardoso, "Estimating equations for source separation," *Proceeedings of the ICASSP-97,* Apr. 21-24, Munich, Germany, 1997.
[8] P. Comon, "Independent component analysis, a new concept?," *Signal Processing,* vol. 36, no. 3, pp. 287-314, Apr. 1994.
[9] M. P. Cooke, *Modelling Auditory Processing and Organisation,* Cambridge University Press, Cambridge, 1993.
[10] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment ," *IEEE Transactions on Signal Processing,* vol. 45, no. 10, Oct. 1997.
[11] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," *Dissertation,* Massachusetts Institute of Technology, Cambridge, 1996.
[12] J. Huang, N. Ohnishi, X. Guo and N. Sugie, "Echo avoidance in a computational model of the precedence effect," *Speech Communication,* vol. 27, pp. 223-233, 1999.
[13] C. Jutten and J. Herault, "Blind separation of sources, Parts I - III," *Signal Processing,* vol. 24, pp. 1-29, 1991.
[14] K. Matsuoka, M. Ohya and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks,* vol. 8, no. 3, pp. 411-419, 1995.
[15] T. Nakatani, T. Kawabata and H. Okuno, "Computational model of sounds stream segregation with multi-agent paradigm," *IEEE ICASSP International Conference on Acoustics, Speech and Signal Processing,* vol. 4, pp. 2671-2674, 1995.
[16] A. Prieto, C. G. Puntonet and B. Prieto, "A neural algorithm for blind separation of sources based on geometric properties," *Signal Processing,* vol. 64, pp. 315-331, 1998.
[17] D. F. Rosenthal and H. G. Okuno, Editors, *Computational Auditory Scene Analysis,* Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

[18] A. J. W. van der Kouwe and D. L. Wang, "Temporal alignment, spatial spread and the linear independence criterion for blind separation of voices," *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Chicago, IL, pp. 1994-1996, 1997.

[19] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks,* vol. 10, no. 3, pp. 684-697, 1999.

[20] M. Weintraub, *A Theory and Computational Model of Monaural Auditory Sounds Separation,* PhD Thesis, Stanford Unviersity, 1985.

[21] H. H. Yang, S. Amari and A. Cichocki, "Information-theoretic approach to blind separation of sources in non-linear mixture," *Signal Processing,* vol. 64, pp. 291-300, 1998.