A Comparison of Auditory and Blind Separation Techniques for Speech Segregation

André J. W. van der Kouwe, *Member, IEEE*, DeLiang Wang, *Member, IEEE*, and 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.**

***Index Terms*—Auditory scene analysis, blind source separation, computational auditory scene analysis (CASA), oscillatory correlation, speech segregation.**

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 organization, 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 [4]. 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) [5], [23]. 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 [16]. 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 [13], [2].

The purpose of this paper is to compare the CASA approach to speech segregation with the blind source separation approach. We choose representative techniques from each domain and apply them 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 techniques, and hence to suggest directions for future research.

To represent CASA we have chosen the system of Wang and Brown [25]. Blind source separation is represented by the second order blind identification (SOBI) algorithm of Belouchrani *et al.* [3] and the joint approximate diagonalization of eigen-matrices (JADE) algorithm of Cardoso *et al.* [9]. The CASA and blind source separation domains are reviewed in Sections II and Section 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 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 [26]. 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, nonperiodic, 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 [5] 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

Manuscript received June 10, 1999; revised May 23, 2000. This work was completed in part while G. J. Brown was a Visiting Scientist with the Center for Cognitive Science, The Ohio State University, Columbus. A. J. van der Kauwe was at the Biomedical Engineering Center, The Ohio State University, and supported by the Section of Neurological Computing at the Cleveland Clinic Foundation. D. Wang was supported in part by an ONR Young Investigator Award and a grant from NUWC.

A. J. W. van der Kouwe is with the Nuclear Magnetic Resonance Center, Massachusetts General Hospital, Charlestown, MA 02129 USA.

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

G. J. Brown is with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP U.K.

Publisher Item Identifier S 1063-6676(01)01491-2.

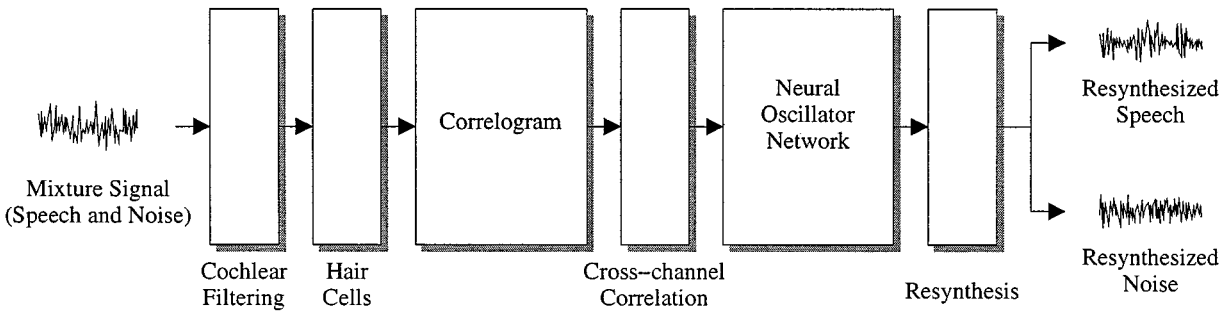

Fig. 1. Schematic diagram of the Wang and Brown CASA model. The input first passes through a model of the auditory periphery (cochlear filtering and hair cells) which simulates auditory nerve activity. Midlevel auditory representations are then formed (correlogram and cross-channel correlation map). 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 [25]).

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 modeled using various auditory maps, which represent periodicities, frequency transitions, onsets 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 [25] 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—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 Fig. 1. Note that the system allows a time-domain waveform to be resynthesized for each segregated sound source. The principal feature that is used for grouping is fundamental frequency (F0); information about the F0 of sound sources is derived from an autocorrelation analysis of each auditory filter channel, forming a representation known as the "correlogram."

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., [14], [21]). 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 SOBI and JADE algorithms.

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. As work in the area proceeds, these requirements are becoming more relaxed [8]. When the 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). The second order blind identification (SOBI) algorithm [3] uses only stationary second order statistics and is based on the joint diagonalization of a set of covariance matrices. Algorithms that operate explicitly on higher-than second order statistics are classified as independent component analysis (ICA) (we note that some nonlinear PCA algorithms implicitly operate on higher order statistics [16]). For example, the ICA procedure of Comon [10] minimizes the fourth order cumulants given by (1) after whitening[1] the signals.

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

where $\overline{y}$ is the vector of approximately separated signals. The joint approximate diagonalization of eigen-matrices (JADE) algorithm [7] minimizes the cumulant given by

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

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

TABLE I
INTRUSIVE SOURCES (FROM [11])

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

It has been shown that the performance of algorithms that implement (1) and (2) are equivalent, but a faster optimization process exists for JADE [9].

The simplest model for the mixing process is [20]

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

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 sensors. 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 [24]. 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 [13]:

$$m_i(t) = \sum_j (h_{ij} * s_j)(t) \tag{4}$$

where h_i is the filter impulse response. Finally, the mixing model may be nonlinear [27]. A number of adaptive algorithms have been described for on-line separation of linear and nonlinear mixtures. One algorithm uses a recursive gradient descent approach to maximize the kurtosis of the separated signals [18]. Amari *et al.*describe an adaptive algorithm for separating multichannel data that incorporates deconvolution and therefore handles the case of convolutive distortion, delays and echoes [1].

The existing blind source separation algorithms vary in terms of three main 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 SOBI and JADE algorithms based on these three properties. Both algorithms assume 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. SOBI is a second order technique and JADE is a fourth order technique. Since the CASA algorithm with which blind separation is compared is autocorrelation-based and therefore relies on second order statistics, it is appropriate to compare it with a second order blind separation technique. Several algorithms are iterative, converging on a set of parameters for the separating model. SOBI and JADE are not iterative, but act on the statistics of the complete set of data directly. For separating linear mixtures of signals, SOBI is representative of second order approaches and JADE represents the best that any equivalent fourth order iterative method can do.

IV. TEST DATA

We use the corpus of sound mixtures devised by Cooke [11] for our evaluations. It consists of 100 speech and noise mixtures, formed by combining each of voiced utterances ($v0$ to $v9$) with each of ten 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. Fully voiced utterances were used because Wang and Brown's CASA system separates acoustic sources according to their F0s. However, this is not a general limitation of CASA systems; for example, Okuno *et al.* [22] describe an approach which uses spatial location cues to segregate two utterances consisting of voiced and unvoiced speech sounds.

For SOBI and JADE, 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. However this is not a general limitation of blind source separation techniques. 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 [6]. It has also been demonstrated that higher-order statistics can be used to recover more sources than mixtures [12], [19], [28].

In this experiment, 100 mixture *pairs* were created by combining the voice and intrusion signals in the ratio 1 : 0.8 and

TABLE II
VOICED SOURCES (FROM [11])

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

0.8 : 1. 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 the mixture consists of exactly two components. The results of this experiment with SOBI and JADE are contrasted with those of the CASA algorithm of Wang and Brown. 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 ten voice signals after conversion to decibels.

Wang and Brown presented their system with only a single mixture of signals, of which the SNR was equal to the average of the two presented to SOBI and JADE. The SNR values for the mixtures before separation are therefore slightly higher for the blind separation techniques (Fig. 2) than for the CASA technique (Fig. 3), because the higher SNR of the two mixtures available in each case (the 1.0 : 0.8 mixture) was chosen for a fair comparison.

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

$$\overline{y}(t) = \mathbf{N}^{-1}\overline{m}(t) \tag{5}$$

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

$$SNR = \left[\frac{m_{00}n_{11} - m_{10}n_{01}}{m_{01}n_{11} - m_{11}n_{01}}\right]^2 \tag{6}$$

where the definitions for the parameters and the derivation are given in the Appendix.

The chosen CASA technique provides a resynthesis path for the separated signals [25], 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.

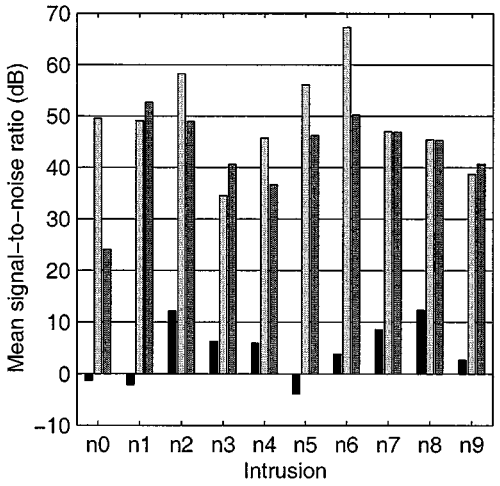

Fig. 2. Comparison of SNR before and after segregation by SOBI and JADE. The left bar denotes initial SNR, the middle bar denotes SNR after segregation by SOBI and the right bar denotes SNR after segregation by JADE. Voiced speech is segregated from a mixture of speech and ten different intrusions (n0 = 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; and n9 = female speech).

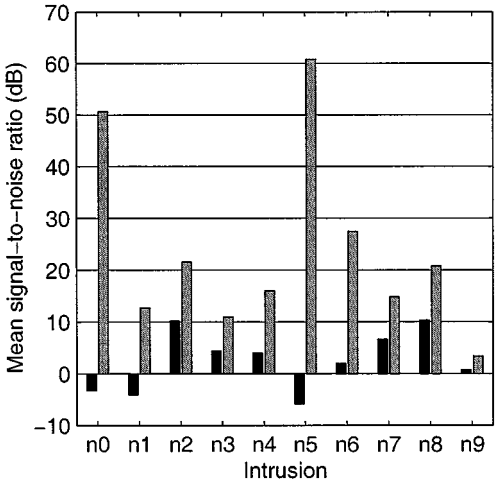

Fig. 3. SNR before and after segregation by computational auditory scene analysis. Voiced speech is segregated from a mixture of speech and ten different intrusions labeled as in Fig. 2 (from [25]).

V. COMPARISONS

The average SNR values (in decibels) for SOBI and JADE are given in Fig. 2. For comparison, the Wang and Brown results [25] are shown in Fig. 3.

In order to compare these results meaningfully, it is necessary to contrast the requirements of SOBI and JADE with those of the CASA technique.

- For SOBI/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 several of these requirements. However, 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

TABLE III
SOURCE SIGNAL PROPERTIES RELEVANT TO SOBI AND JADE

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

well-defined regions corresponding to one or more of the source signals.

The CASA approach imposes less stringent requirements on the statistical and mixing properties of the signals than the SOBI and JADE algorithms. The most important of these are properties 6 and 7 in Table III—SOBI and JADE require 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, 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. It should be noted that blind separation algorithms have been described to deal with the situation where the number of sources is unequal to and possibly greater than the number of mixtures [12], [19], [28] and the problem of determining the unknown number of sources has been addressed [3], [28].

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 [24]. Some blind separation algorithms handle distortions and delays by modeling convolutive processes [2]. Amari *et al.* describe an on-line adaptive algorithm for blind deconvolution, generalized for the case where the number of sensors may be more or less than the number of sources [1]. Zhang *et al.* present a natural gradient approach that also deals with over- and under-complete mixtures [28]. Signals containing echoes have not been dealt with by CASA algorithms, but one would expect a lesser problem because segregation in CASA algorithms is 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 [15] could be incorporated into CASA systems in order to minimize the distortion caused by echoes.
- 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. On-line adaptive algorithms that are able to handle mixing processes that vary with time have been described. For example, LeBlanc and De Leòn describe an adaptive algorithm that uses kurtosis maximization to separate voice signals [18]. 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 [20], *require* that the signals be nonstationary and may be able to handle a mixing process that varies slowly with time.

The results of our tests, shown in Figs. 2 and 3, reveal the following:

- JADE performs best for the white noise $(n1)$, and it performs worst for the 1 kHz tone $(n0)$;
- SOBI performs similarly to JADE for the voice signals $(n7, n8$ and $n9)$, somewhat better for the noise bursts $(n2)$ and rock music $(n4)$, and significantly better for the 1 kHz tone $(n0)$ and the trill telephone $(n6)$;
- 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 SOBI it is the second order statistics that are used, and in JADE the statistics up to the fourth order are used. 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 the performance differences. Hence, the poor performance of JADE on the 1 kHz tone intrusion may be explained by the fact that the tone mixtures yield poor higher order statistics. Similarly, the white noise mixtures contain rich higher order joint statistics, which can be exploited by JADE. SOBI is likely to perform well in situations where there is good spectral separation between sources. This accounts for its relatively high performance on the $n0$ and $n6$ mixtures, since the tone and telephone both have their energy concentrated in narrow spectral regions.

The greatest contributor to the variation in performance of the CASA technique over the range of intrustions is the structure of the intrusive sound in the time-frequency plane. This is characterized 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 (and to a lesser extent, the trill telephone). 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 F0 tracking procedure.

Finally, it may be noted that there are limits on the ability of human listeners to separate simultaneous events. This is well demonstrated by the 'cocktail party effect' [4]. In such listening situations, humans are unable to distinguish every conversation; however, they have a remarkable ability to attend selectively to the voice of a single speaker. Wang and Brown's CASA algorithm works in a similar manner; it separates a target speech signal (the "foreground") from interfering noise (the "background"). In contrast, blind separation techniques typically attempt to segregate every source signal from a mixture.

VI. Conclusion

Blind separation can be remarkably successful at separating mixtures of sounds, provided that certain requirements on the properties of the source signals are met. These requirements may not be equitable with a natural listening environment. As work in blind separation progresses, the requirements are becoming ever less restrictive. If the 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.

In most of the noise conditions used here, the blind separation techniques outperform CASA by some degree. Cooke's set of acoustic mixtures [11] was designed to present a challenging test for CASA systems, and it does so by including a wide range of possible noise intrusions. However, the corpus uses a simple linear mixing model which satisfies all of the assumptions required by JADE and SOBI. Similarly, other corpora intended for testing CASA, such as the multi-speaker corpus described by Karlsen *et al.* [17] lack the sensor array recordings required for blind separation algorithms. Clearly, further comparison of CASA and blind separation approaches would be facilitated by the use of a common corpus which was designed with both techniques in mind.

Finally, the different performance profiles of the CASA and blind separation techniques suggest that there would be merit in combining the two approaches. More specifically, scene analysis heuristics that are employed by CASA systems (such as continuity of F0 and spatial location), could be exploited by blind separation algorithms in order to improve their performance on real-world acoustic mixtures. Conversely, blind separation techniques could help CASA in decomposing mixtures that overlap substantially in the time-frequency plane.

Appendix
Derivation of (6)

If the components of the vectors and matrices in (3) are defined as $\overline{\mathbf{M}} = \begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix}$, $\overline{m}(t) = \begin{pmatrix} m_0(t) \\ m_1(t) \end{pmatrix}$ and $\overline{s}(t) = \begin{pmatrix} s_0(t) \\ s_1(t) \end{pmatrix}$ then (3) may be expressed in scalar form as follows:

$$m_0 = m_{00}s_0 + m_{01}s_1 \tag{A1}$$

$$m_1 = m_{10}s_0 + m_{11}s_1. \tag{A2}$$

If the components of the vectors and matrices in (5) are defined as $\overline{\mathbf{N}} = \begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix}$ and $\overline{y}(t) = \begin{pmatrix} y_0(t) \\ y_1(t) \end{pmatrix}$ then rearranging the scalar form of (5) yields

$$y_0 = \frac{m_0}{n_{00}} - \frac{n_{01}}{n_{00}}y_1 \tag{A3}$$

$$y_1 = \frac{m_1}{n_{11}} - \frac{n_{10}}{n_{11}}y_0. \tag{A4}$$

Combining (A3) and (A4) gives

$$y_1 = \frac{m_{00}n_{11} - m_{10}n_{01}}{n_{00}n_{11} - n_{01}n_{10}}. \tag{A5}$$

Combining (A1), (A2), and (A5) gives

$$y_0 = \left(\frac{m_{00}n_{11} - m_{10}n_{01}}{n_{00}n_{11} - n_{01}n_{10}}\right)s_0 + \left(\frac{m_{01}n_{11} - m_{11}n_{01}}{n_{00}n_{11} - n_{01}n_{10}}\right)s_1. \tag{A6}$$

If s_0 is chosen to be the original signal and s_1 the original noise, then y_0 is the estimate of the signal after separation. It then follows that the SNR of y_0 is

$$SNR_{y_0} = \left(\frac{m_{00}n_{11} - m_{10}n_{01}}{m_{01}n_{11} - m_{11}n_{01}}\right)^2 \frac{s_0^2}{s_1^2} \tag{A7}$$

from which follows (6).

Acknowledgment

The authors thank the three anonymous reviewers whose constructive comments improved the presentation of this paper.

References

[1] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," in *Proc. 11th IFAC Symp. System Identification*, Kitakyushu City, Japan, July 1997, pp. 1007–1012.

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 434–444, Feb. 1997.

[4] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[5] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[6] X. Cao and R. Liu, "General approach to blind source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 562–571, Mar. 1996.

[7] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for nongaussian signals," *Proc. Inst. Elect. Eng.*, vol. 140, pp. 362–370, Dec. 1993.
[8] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.
[9] ——, "High-order contrasts for independent component analysis," *Neural Comput.*, vol. 11, pp. 157–192, Jan. 1999.
[10] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, Apr. 1994.
[11] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[12] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, "ICA algorithms for 3 sources and 2 sensors," in *Proc. IEEE Workshop Higher-Order Statistics*, Caesarea, Israel, June 1999, pp. 116–120.
[13] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. Signal Processing*, vol. 45, pp. 2068–2612, Oct. 1997.
[14] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. thesis, Mass. Inst. Technol., Cambridge, MA, 1996.
[15] J. Huang, N. Ohnishi, X. Guo, and N. Sugie, "Echo avoidance in a computational model of the precedence effect," *Speech Commun.*, vol. 27, pp. 223–233, 1999.
[16] C. Jutten and J. Herault, "Blind separation of sources, parts I–III," *Signal Process.*, vol. 24, pp. 1–29, 1991.
[17] B. L. Karlsen, G. J. Brown, M. P. Cooke, M. D. Crawford, P. D. Green, and S. J. Renals, "Analysis of a multi-simultaneous-speaker corpus," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum, 1998, pp. 321–333.
[18] J. P. LeBlanc and P. L. De Leòn, "Speech separation by kurtosis maximization," in *Proc. 1998 IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Seattle, WA, 1998, pp. 1029–1032.
[19] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Lett.*, vol. 6, pp. 87–90, Apr. 1999.
[20] 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.
[21] T. Nakatani, T. Kawabata, and H. Okuno, "Computational model of sounds stream segregation with multi-agent paradigm," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, 1995, pp. 2671–2674.
[22] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Commun.*, vol. 27, pp. 299–310, 1999.
[23] D. F. Rosenthal and H. G. Okuno, Eds., *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum, 1998.
[24] A. J. W. van der Kouwe and D. L. Wang, "Temporal alignment, spatial spread and the linear independence criterion for blind separation of voices," in *Proc. 19th Annu. Int. Conf. IEEE Engineering Medicine Biology Society*, Chicago, IL, 1997, pp. 1994–1996.
[25] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, May 1999.
[26] M. Weintraub, "A Theory and Computational Model of Monaural Auditory Sounds Separation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1985.
[27] H. H. Yang, S. Amari, and A. Cichocki, "Information-theoretic approach to blind separation of sources in nonlinear mixture," *Signal Process.*, vol. 64, pp. 291–300, 1998.
[28] L.-Q. Zhang, S. Amari, and A. Cichocki, "Natural gradient approach to blind separation of over- and under-complete mixtures," in *Proc. Int. Workshop Independent Component Analysis Blind Signal Separation*, Aussois, France, Jan. 1999, pp. 455–460.

André J. van der Kouwe (M'90) received the B. Eng. and M. Eng. degrees in electrical and electronic engineering from the University of Pretoria, Pretoria, South Africa, in 1992 and 1995, respectively, and the Ph.D. degree in biomedical engineering from The Ohio State University, Columbus, in 1999. His doctoral research was done in the Department of Neurology, Cleveland Clinic Foundation.

He is currently a research fellow at the Nuclear Magnetic Resonance Center, Massachusetts General Hospital—Harvard Medical School, Charlestown, MA. His research interests include MRI pulse sequence optimization, and the clinical applications of signal/image processing in imaging and electrophysiological monitoring of the human brain.

DeLiang Wang (M'94) received the B.S. degree in 1983 and the M.S. degree in 1986 from Peking University, Beijing, China, and the Ph.D. degree in 1991 from the University of Southern California, Los Angeles, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer and Information Science and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently an Associate Professor. From October 1998 to September 1999, he was a Visiting Scholar with the Vision Sciences Laboratory, Harvard University, Cambridge, MA. His present research interests include neural networks for perception, neurodynamics, neuroengineering, and computational neuroscience.

Dr. Wang is a recipient of the 1996 U.S. Office of Naval Research Young Investigator Award.

Guy J. Brown received the B.Sc. degree in applied science from Sheffield Hallam University, U.K., in 1988, the Ph.D. degree in computer science and the M.Ed. degree from the University of Sheffield, Sheffield, U.K., in 1992 and 1997, respectively.

He is currently a Senior Lecturer in computer science with the University of Sheffield. He has studied computational models of auditory perception since 1989, and also has research interests in speech perception, computer-assisted learning, and music technology.