# Dear Author

Here are the proofs of your article.

- You can submit your corrections **online** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- Please return your proof together with the permission to publish confirmation.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the journal title, article number, and your name when sending your response via e-mail, fax or regular mail.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.

**Please note**

Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

`http://dx.doi.org/10.1007/s11265-009-0443-3`

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

The **printed version** will follow in a forthcoming issue.

**To:** **Springer Customer Support 2**
**E-mail: CorrAdmin2@spi-bpo.com**
**Fax: +1-703-5621873**
**SPi SPi Building, Sacsac Bacong Oriental Negros 6216 Philippines**

**Re:**

**Journal of Signal Processing Systems DOI 10.1007/s11265-009-0443-3**

Correlated Postfiltering and Mutual Information in Pseudoanechoic Model Based Blind Source Separation

**Persia · Milone · Yanagida**

## Permission to publish

I have checked the proofs of my article and

- ❑ I have **no corrections.** The article is ready to be published without changes.
- ❑ I have **a few corrections.** I am enclosing the following pages:
- ❑ I have made **many corrections.** Enclosed is the **complete article.**

**Date / signature:** ______________________________________________

# Metadata of the article that will be visualized in Online

**Please note: Image will appear in color online but will be printed in black and white.**

| | | | |
|---|---|---|---|
| 1 | Article Title | **Correlated Postfiltering and Mutual Information in Pseudoanechoic Model Based Blind Source Separation** | |
| 2 | Article Sub- Title | | |
| 3 | Article Copyright - Year | **Springer Science+Business Media, LLC. Manufactured in The United States 2009<br>(This will be the copyright line in the final PDF)** | |
| 4 | Journal Name | Journal of Signal Processing Systems | |
| 5 | Corresponding Author | Family Name | **Persia** |
| 6 | | Particle | **Di** |
| 7 | | Given Name | **Leandro Ezequiel** |
| 8 | | Suffix | |
| 9 | | Organization | Universidad Nacional del Litoral |
| 10 | | Division | Facultad de Ingeniería y Ciencias Hídricas |
| 11 | | Address | Ciudad Universitaria, Paraje “El Pozo”, Santa Fe S3000, Argentina |
| 12 | | e-mail | ldipersia@gmail.com |
| 13 | Author | Family Name | **Milone** |
| 14 | | Particle | |
| 15 | | Given Name | **Diego H.** |
| 16 | | Suffix | |
| 17 | | Organization | Universidad Nacional del Litoral |
| 18 | | Division | Facultad de Ingeniería y Ciencias Hídricas |
| 19 | | Address | Ciudad Universitaria, Paraje “El Pozo”, Santa Fe S3000, Argentina |
| 20 | | e-mail | dmilone@fich.unl.edu.ar |
| 21 | Author | Family Name | **Yanagida** |
| 22 | | Particle | |
| 23 | | Given Name | **Masuzo** |
| 24 | | Suffix | |
| 25 | | Organization | Doshisha University |
| 26 | | Division | Department of Intelligent Information Engineering and Science |
| 27 | | Address | Kyotanabe, Kyoto 610-0321, Japan |
| 28 | | e-mail | myanagid@mail.doshisha.ac.jp |
| 29 | Schedule | Received | 21 April 2009 |
| 30 | | Revised | 14 August 2009 |
| 31 | | Accepted | 4 December 2009 |
| 32 | Abstract | In a recent publication the pseudoanechoic mixing model for closely spaced | |

microphones was proposed and a blind audio sources separation algorithm based on this model was developed. This method uses frequency-domain independent component analysis to identify the mixing parameters. These parameters are used to synthesize the separation matrices, and then a time-frequency Wiener postfilter to improve the separation is applied. In this contribution, key aspects of the separation algorithm are optimized with two novel methods. A deeper analysis of the working principles of the Wiener postfilter is presented, which gives an insight in its reverberation reduction capabilities. Also a variation of this postfilter to improve the performance using the information of previous frames is introduced. The basic method uses a fixed central frequency bin for the estimation of the mixture parameters. In this contribution an automatic selection of the central bin, based in the information of the separability of the sources, is introduced. The improvements obtained through these methods are evaluated in an automatic speech recognition task and with the PESQ objective quality measure. The results show an increased robustness and stability of the proposed method, enhancing the separation quality and improving the speech recognition rate of an automatic speech recognition system.

| | | |
|---|---|---|
| 33 | Keywords separated by ' - ' | Pseudoanechoic model - Blind source separation - Automatic speech recognition - Mutual information - Wiener postfilter |
| 34 | Foot note information | This work was supported by ANPCYT under projects PICT 12700 and PICT 25984, CONICET, and UNL under project CAI+D 012-72. |

J Sign Process Syst
DOI 10.1007/s11265-009-0443-3

# Correlated Postfiltering and Mutual Information in Pseudoanechoic Model Based Blind Source Separation

Q1 **Leandro Ezequiel Di Persia · Diego H. Milone · Masuzo Yanagida**

Received: 21 April 2009 / Revised: 14 August 2009 / Accepted: 4 December 2009




**Abstract** In a recent publication the pseudoanechoic mixing model for closely spaced microphones was proposed and a blind audio sources separation algorithm based on this model was developed. This method uses frequency-domain independent component analysis to identify the mixing parameters. These parameters are used to synthesize the separation matrices, and then a time-frequency Wiener postfilter to improve the separation is applied. In this contribution, key aspects of the separation algorithm are optimized with two novel methods. A deeper analysis of the working principles of the Wiener postfilter is presented, which gives an insight in its reverberation reduction capabilities. Also a variation of this postfilter to improve the performance using the information of previous frames is introduced. The basic method uses a fixed central frequency bin for the estimation of the mixture parameters. In this contribution an automatic selection of the central bin, based in the information of the separability of the sources, is introduced. The improvements obtained through these methods are evaluated in an automatic speech recognition task and with the PESQ objective quality measure. The results show an increased robustness and stability of the proposed method, enhancing the separation quality and improving the speech recognition rate of an automatic speech recognition system.

**Keywords** Pseudoanechoic model · Blind source separation · Automatic speech recognition · Mutual information · Wiener postfilter

This work was supported by ANPCYT under projects PICT 12700 and PICT 25984, CONICET, and UNL under project CAI+D 012-72.

Q2 L. E. Di Persia (✉) · D. H. Milone
Facultad de Ingeniería y Ciencias Hídricas,
Universidad Nacional del Litoral, Ciudad Universitaria,
Paraje "El Pozo", Santa Fe, S3000, Argentina
e-mail: ldipersia@fich.unl.edu.ar, ldipersia@gmail.com

D. H. Milone
e-mail: dmilone@fich.unl.edu.ar

M. Yanagida
Department of Intelligent Information
Engineering and Science, Doshisha University,
Kyotanabe, Kyoto, 610-0321, Japan
e-mail: myanagid@mail.doshisha.ac.jp

## 1 Introduction

One of the fundamental problems for the widespread of applications of automatic speech recognition is the degrading effect of noise [14]. The speech recognition systems trained under laboratory conditions, suffer a strong degradation in their performance when used in real environments [20]. Several aspects contribute to this degrading effect. One of them is the presence of multiple sound sources other than the desired one, which alter the information of the desired source and produce a deterioration of the recognition rate. Another problem is related to the use of distant microphones [18]. In an ideal close talking environment the microphones used to capture the sound field are located near to the speaker mouth. In this way, the direct sound from the target speech is picked with a large signal to noise ratio. But in several applications, like teleconference systems or remote controlling of home appliances, the microphones are located far away from the speaker. In this way the sound field that the microphones pick up is affected by several sound sources in a stronger way, producing a lower SNR. Moreover, the target speech is modified by the room impulse response, producing a smearing in its contents and a coloring of the spectra [12]. This effect is known as reverberation, and it affects the performance of ASR systems even if there are no other sound sources and if the system was trained with speech recorded in the same conditions [2].

There are several approaches that try to mitigate the competing noise effect. Basically the alternatives are applied at three different levels of the speech recognition system [10]. At the level of the audio signal, the enhancement approach tries to produce a speech signal as similar to the original source as possible. At the level of the features used by the recognizer, the robustness is introduced either by using a set of intrinsically robust features, or by projecting the noisy features on the space of clean features. Finally, at the level of the acoustic models, the effect of noise can be introduced either by using multiple acoustic models for different noise conditions, or by an adaptation of the basic model to the noise conditions during the use of the system. This work is focused in the first kind of techniques, the task is to preprocess the audio signal to produce a desired speech signal as clean as possible. In particular, this is done using multiple input signals captured through a microphone array.

In particular this work is focused in a recently proposed frequency-domain independent component analysis (fd-ICA) algorithm, which uses a pseudo-anechoic mixing model, under the assumption of closely spaced microphones. This separation method, named pseudoanechoic model blind source separation (PMBSS) was shown to be very effective in produce separation in environments where other approaches fail, and with a very high processing speed [8]. For example, it can produce an improvement of more than a 45% in recognition rate, with a processing speed more than 16 times higher than the standard method proposed by Parra et al. [19].

This contribution will be focused in producing relevant improvements to the PMBSS method. First, a revision of the PMBSS method will be presented, including a new analysis of the working principles of the Wiener postfilter, that show its capabilities to not only enhance the separation, but also of reducing the reverberation. Next, two alternative methods will be presented, one for automatic selection of the optimal central frequency to use in the estimation of the mixing parameters, and the other in the Wiener postfilter, to exploit the temporal information in the noise estimation. This section will be followed by a series of experiments to show improvements introduced by the proposed methods. A discussion and conclusion section ends the article.

## 2 Pseudoanechoic Model for BSS

In this work the speech enhancement approach is used. In this way the objective will be to obtain a speech as clean as possible. Among the many techniques used for this purpose, the microphone array processing has recently received strong attention from the scientific community. The task of blind source separation in the microphone array context, consist in the extraction of the sources that originated the sound field, given a set of measurements obtained through an array of microphones [12].

The problem is known in the literature as "cocktail party", because of the analogy with such a party in which there are several speakers and sound sources, and yet human beings have the ability to segregate the source of interest and concentrate in the desired conversation [11]. This ability is related to the fact that humans have two ears, and thus a multi-microphone setup is naturally introduced as an alternative for the solution. A brief mathematical description of the problem will be presented in the following.

### 2.1 Convolutive BSS Problem

Consider the case in which there are $M$ active sound sources, and the sound field generated by them is captured by $N$ microphones, as shown in Fig. 1. From source $j$ to microphone $i$, an impulse response $h_{ij}$ characterizes the room. Using the notation $s_j$ for the sources and $x_i$ for the microphone signals, with $i = 1, \ldots, N$ and

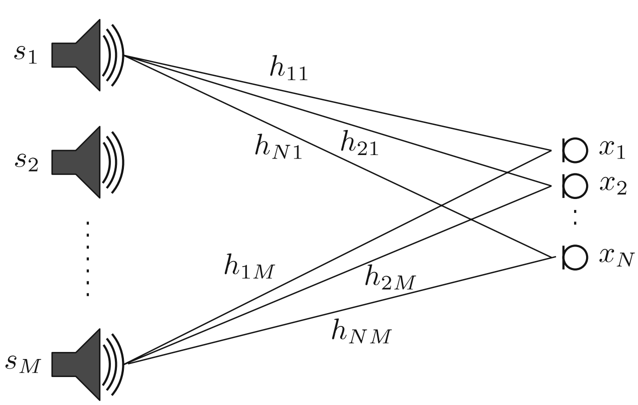


**Figure 1** A case of cocktail party with $M$ sources and $N$ microphones.

$j = 1, \ldots, M$, the mixture can be represented at each instant $t$ as [4]:

$$x_i(t) = \sum_j h_{ij}(t) * s_j(t) \ , \quad (1)$$

where $*$ stands for convolution. Let us form a vector of sources,

$$\mathbf{s}(t) = \left[s_1(t), \cdots, s_M(t)\right]^T$$

and the same for the vector of mixtures

$$\mathbf{x}(t) = \left[x_1(t), \cdots, x_N(t)\right]^T$$

measured by the microphones, where $[\cdot]^T$ stands for transposition. Then the previous equation can be written (with a little abuse of notation) as:

$$\mathbf{x}(t) = H * \mathbf{s}(t) \quad (2)$$

where the "matrix" $H$ has as each element a filter given by the impulse response from one source location to one microphone location. The equation must be understood as a simple matrix-vector product, but replacing the multiplications by a filtering operation via convolution.

In this context, there are several approaches for the solution of the BSS problem. From the basic ones based on beamforming [3], to the more advanced separation methods based the sparcity of the sources in the time-frequency domain [25] and the separation based on the search of statistical independence of the obtained sources [9]. The last approach assumes that the original sources are statistically independent, and thus the separation can be achieved searching for a transformation that produces statistically independent results. This approach uses independent component analysis (ICA) and there are several methods that exploit the independence to yield the estimated sources.

One of the more successful methods is the frequency-domain independent component analysis method (fd-ICA) [23]. If a short time Fourier transform (STFT) is applied to Eq. 2, the mixture can be written as [2, Chapter 13]

$$\mathbf{x}(\omega, \tau) = H(\omega)\, \mathbf{s}(\omega, \tau) \ , \quad (3)$$

where the variable $\tau$ represents the time localization given by the sliding window in the STFT, and $\omega$ is the frequency. It should be noted that, as the mixing system was assumed to be LTI, the matrix $H(\omega)$ is not a function of the time. Also note that the convolution operations have been replaced by ordinary multiplication, which makes the problem simpler in this domain.

The classical solution alternative is to apply an ICA algorithm to each frequency bin, producing separation on each of them. After separation, the separated sources in each bin need to be reordered due to the permutation ambiguity inherent to ICA methods, and then an inverse STFT is used for the time-domain reconstruction. The permutation problem is one of the main drawbacks of this method, because its correction is not trivial, and although many solution alternatives have been proposed, none of them is completely effective [17]. Another problem of the standard method is the different convergence of the ICA method for each frequency bin, which yields different separation qualities for different bins, including some bins where the method failed to converge to a proper solution.

### 2.2 The Pseudoanechoic Model

In a previous development [8], the pseudoanechoic model was proposed as an alternative to solve this problem. If the microphones are closely spaced, it can be assumed that the impulse response from a source to all the microphones will be delayed and scaled versions of it. Using the notation of Fig. 1, with $M = N = 2$, the mixture can be expressed as

$$x_1(t) = s_1(t) * h_{11}(t) + s_2(t) * h_{12}(t)$$
$$x_2(t) = s_1(t) * h_{21}(t) + s_2(t) * h_{22}(t) \ . \quad (4)$$

Under the assumption of closely spaced microphones, the crossing impulse response can be expressed as delayed and scaled version of the direct impulse responses, approximating $h_{21}(t) \simeq \alpha h_{11}(t - d_1)$ and $h_{12}(t) \simeq \beta h_{22}(t - d_2)$. This simplification is important because it allows to write the mixing matrix of Eq. 3 in a simpler way

$$\mathbf{x}(\omega, \tau) = \begin{bmatrix} 1 & \beta e^{-jd_2\omega} \\ \alpha e^{-jd_1\omega} & 1 \end{bmatrix} \begin{bmatrix} H_{11}(\omega) & 0 \\ 0 & H_{22}(\omega) \end{bmatrix} \mathbf{s}(\omega, \tau) \quad (5)$$

In this equation, the rightmost matrix, which does not produces any mixing, represent the room effect on each source signal. The leftmost matrix in turn, represents the mixing effect. In this way the very complex filtering and mixing effect of the room can be decomposed in two simpler parts, one of mixing and the other of filtering. Applying the filtering part to the source signals, the following is obtained

$$\mathbf{x}(\omega, \tau) = \begin{bmatrix} 1 & \beta e^{-jd_2\omega} \\ \alpha e^{-jd_1\omega} & 1 \end{bmatrix} \mathbf{z}(\omega, \tau) \quad (6)$$

where now the $\mathbf{z}(\omega, \tau)$ contains the reverberated sources. In simple words, the pseudoanechoic model concentrate the effect of the room in a general impulse response for each channel which introduces distortion to that signal, and a simpler mixing which is similar to the anechoic model which is applied on these reverberant signals. It was shown that this model is plausible for microphones separated even by 5 cm, in moderate reverberant conditions.

Based on this mixing model, the PMBSS algorithm was introduced. Simply speaking, this method aims to produce the $\mathbf{z}$ sources mentioned before. It is interesting to note that in Eq. 6, the mixing matrix has a dependency on $\omega$ which is easy to synthesize. For all frequencies, the parameters $\alpha$, $\beta$, $d_1$ and $d_2$ have constant values, this means that if one is capable of identifying these parameters in a robust way for one specific frequency, they can be used to synthesize the mixing matrix (and by inversion, the separation matrix) for all the frequencies. Basically, the PMBSS method has three stages: 1) Estimation of the Mixing parameters for a *given* frequency bin, using ICA; 2) Synthesis of the separation matrixes for all frequencies using the estimated parameters, and separation; 3) Application of a time-frequency Wiener postfilter. The main advantage of this method is that instead of performing one ICA separation for each frequency bin, only *one* ICA problem is solved over the data from a given central bin and a small number of lateral bins. From the estimated mixing matrix, the mixing parameters of the pseudoanechoic model are estimated, and used to synthesize the separation matrices for all the bins. In this way the resulting algorithm is extremely fast, and yet it produces a high quality of separation.

The key aspect of this method is how to identify the mixing parameters accurately. The proposed method consisted in using ICA in a previously selected (fixed) frequency bin. Moreover, to produce robustness, instead of the data of only that bin, the data from a group of bins, taken symmetrically around the selected frequency, was used. In this way the ICA algorithm has a lot of data for the learning of the parameters, which can speed up the convergence, and moreover, the estimation produced is more robust, as shown in the previous work. Nevertheless, the selection of the optimal central bin to use was not explored. There must exist an specific frequency bin for which the parameters can be estimated more accurately. If this bin can be identified by an easy method, it can improve the separation results

Another interesting aspect of this method was the introduction of a time-frequency Wiener filter estimated using the information obtained after the separation stage. At this point, an estimation of the reverberant sources $\mathbf{z}(\omega,\tau) = [z_1(\omega,\tau)\ z_2(\omega,\tau)]$ was obtained. As the separation method is not perfect and the main hypothesis may be only partially fulfilled, the separated sources will have some residual components of the competing source. This is because the separation matrix can only reject the source coming from one direction, as shown in [1]. Nevertheless, as the estimations for the two sources are available, this means that to improve the separation of one of the sources, the other can be used as an estimation of the noise. In this way, the time-frequency Wiener filter to improve the source $z_1$ using $z_2$ as an estimation of the noise is given by

$$F_{\mathscr{W},1}(\omega,\tau) = \frac{|z_1(\omega,\tau)|^2}{|z_1(\omega,\tau)|^2 + |z_2(\omega,\tau)|^2}\,, \tag{7}$$

with an equivalent definition for the filter to enhance the other source.

This postfilter was shown to produce an important increase in the separation quality, and also it was shown to be a better alternative than other approaches like binary masks. Nevertheless, the wiener postfilter is a very simple case, and more interesting approaches can be used.

### 2.3 Reverberation Reduction by Wiener Postfilter

In this section a deeper analysis of the Wiener postfilter in a 2 by 2 case is performed, to show how this filtering provides additional reduction, not only of the competing source, but of the echoes coming both from the competing source and the echoes of the desired source. To this end, it is necessary to study the beampatterns generated by the separation matrix. As was shown in [1], the separation matrix generated by ICA works as an adaptive null beamformer, that is, a beamformer which is designed to reject the signal arriving to the microphone array from certain direction. In the two by two case, the separation matrix works as a pair of null beamformers, where each beamformer reject the signals arriving from the estimated direction of arrival of each source.

In an environment with no reverberation, if one of the main signals is eliminated, the resulting signal will have information only of the other signal, and thus producing a good separation. But in reverberant environments, there are echoes arriving to the array from other directions than the main propagation path. As the separation can only eliminate the signal from the main direction, the echoes from both, the desired source and the competing source, will remain in the separated signal.

An uniform linear array of $N$ microphones in the far field is characterized by its array response vector, which is a function of the frequency $f$ and the angle of arrival $\phi$, given by

$$\mathbf{v}(f,\phi) = \left[1, e^{\frac{-j2\pi f d \sin(\phi)}{c}}, e^{\frac{-j2\pi f 2d \sin(\phi)}{c}}, \cdots, e^{\frac{-j2\pi f (N-1) d \sin(\phi)}{c}}\right]^T, \tag{8}$$

where $d$ is the microphone spacing and $c$ the sound speed. This array response vector characterises the microphone array as it explain the relation among the outputs of each of the microphones. If the outputs of the array are linearly combined (as in a delay and sum beamformer), weighted with coefficients $\mathbf{a} = [a_1, a_2, \ldots, a_N]^T$, then the beamformer response $r(f,\phi)$ will be given by

$$r(f,\phi) = \mathbf{a}^H \mathbf{v}(f,\phi) \tag{9}$$

where $[\cdot]^H$ is the conjugate transposed operation. The magnitude of the beamformer response is the array gain or beampattern, which shows for each frequency, how the magnitude of the output signal change with the angle of arrival of the input signals. In the case of the separation matrix, each row of it works as a null beamformer, and thus in a 2 by 2 case a pair of null beamformers is generated. Figure 2 shows the beampatterns generated by the PMBSS method for the case of two speech sources at $\pm 26$ degrees, sampled at 8000 Hz, captured with two microphones spaced by 5 cm. For each beampattern the null is located in the direction of one of the sources.

To analyze the capabilities of this Wiener filter, assume that there is a sound field produced by white and stationary signals, with equal power from all directions. That is, suppose that the microphone array receives equal power from all angles and for all frequencies and times. In this case, the behaviour of the combined separation and Wiener filter process can be analyzed using the beampatterns, as the beampattern output will be the actual magnitude at the output of the separation, as a function of the arrival angle.

Figure 3 shows the beampatterns obtained from the separation matrix in the bin corresponding to 2000 Hz in the same example of Fig. 2 (for other frequencies the analysis is equivalent). The top row shows the beampatterns obtained from the separation matrix. For each beampattern, it can be seen that in the direction of each source, the gain is unitary (which is a consequence of the minimal distortion principle), and in the direction of the other source the gain tends to zero. In the bottom row, we have applied the equation of the Wiener filter to these patterns. That is, if the beamformer gains for the separation matrix at the given frequency are called $G_1(\theta)$ and $G_2(\theta)$, and as they are also the output amplitudes as a function of the angle, the first Wiener filter will be $G_1(\theta)^2/(G_1(\theta)^2 + G_2(\theta)^2)$, and the same for the other filter.

This is a way to visualize the approximate global effect of the whole processing. As it can be seen, the Wiener filter maintains unitary gain in the desired directions and nulls in the interference directions, but also produces attenuation in all other directions, which mitigates the effect of all echoes including both, those from the undesired noise (which improves separation) and these from the desired source (which reduces the reverberation). This is very important, because it means



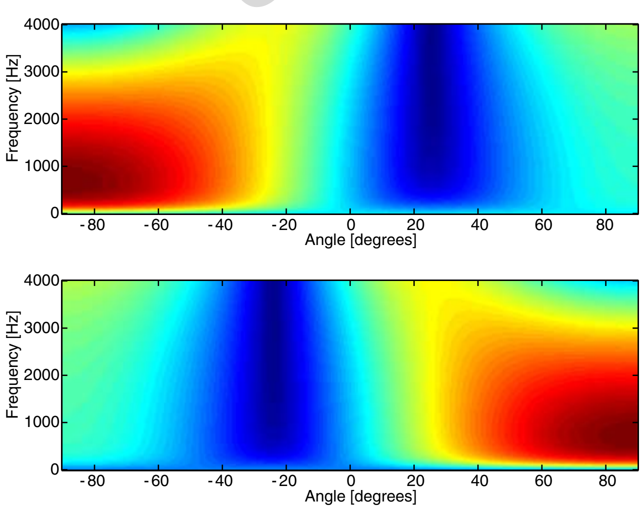


**Figure 2** Beampatterns generated by PMBSS for sources at $\pm 26$ degrees.

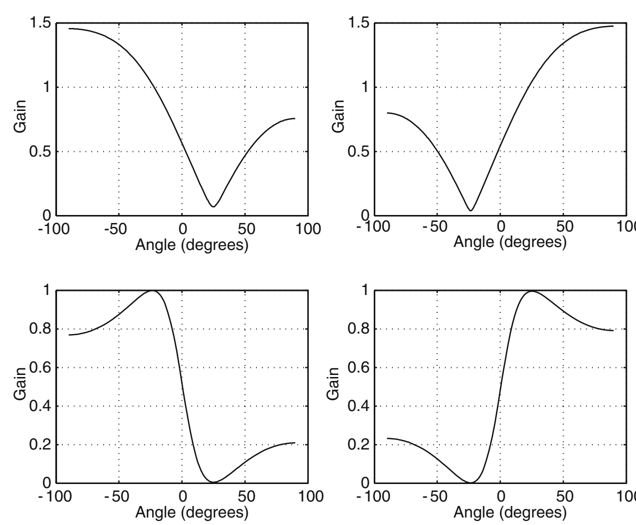


**Figure 3** Effect of the Wiener postfilter on the beampatterns. **a**) the beampatterns generated from the separation matrix. **b**) the beampatterns after application of the Wiener filter.

that it helps in improving the fundamental limitation of the fd-ICA approach as analyzed in [1], that is, the impossibility of rejecting or reducing the echoes. It must be noted that this kind of postfilter is general and can be incorporated in any fd-ICA approach to improve its performance.

Clearly, in real situations the input signals will be neither of the same power for all directions as assumed, nor white and stationary. Nevertheless, the signal with stronger component will in general come from the detected directions, with the echoes of lower power arriving from different directions, and thus the resulting effect would be even better than the depicted one. That is, Fig. 3 represents the worst case of possible inputs, and thus for more realistic cases an even better behaviour can be expected.

## 3 Proposed Methods

As already explained, two improvements for the standard PMBSS method will be introduced. First a method for automatic selection of the central frequency bin to use in the ICA based mixing parameter estimation is introduced. The mutual information provides an estimation of the amount of mixing in each bin. In this way, the selection of a bin which has little overlapping of information will be optimal to find the proper separation.

In second place, the basic time-frequency Wiener postfilter uses an instantaneous time-frequency estimation of the source and noise. But it is know that, due to the reverberation effect, the information in some instant depends also on previous information. To take this effect into account, the noise estimation is composed not only by the present instant but by a number of delayed versions of the previous information. These methods will be introduced in what follows.

### 3.1 Automatic Selection of the Central Bin

As already mentioned, the first stage of PMBSS (estimation of the mixing parameters) is performed by means of a robust ICA method on data collected from a set of frequency centered in a previously chosen bin. In [8], this central bin was set at a fixed value in an arbitrary way. However, for each particular mixture of signals it must be a frequency bin which yields the best possible estimation of the mixing parameters. This optimal bin will depend in the particular sources and mixing characteristics, and thus it would be desirable to have some automatic selection method for it.

The best central bin would be that in which the ICA algorithm can produce the best mixing matrix estimation. Intuitively, it would be one in which, given the characteristics of the mixture, the sources are “less mixed”, or more statistically independent. What is necessary is a measure of how mixed are the signals in each bin. One measure that can be used for this purpose is the mutual information. Mutual information measures the amount of information that is shared among random variables. It is calculated as [5]

$$I(X,Y) = \iint p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy\,, \qquad (10)$$

where $I(X,Y)$ is the mutual information of the two random variables $X$ and $Y$, $p(x,y)$ is the joint probability density function (pdf) of the variables, and $p(x)$ and $p(y)$ are the marginal pdf of the variables. Using the definition of differential entropy $H(X) = -\int p(x)\log(p(x))dx$ and joint differential entropy $H(X,Y) = -\iint p(x,y)\log(p(x,y))dxdy$, the mutual information can be written as [15]

$$I(X,Y) = H(X) + H(Y) - H(X,Y)\,. \qquad (11)$$

The mutual information is always positive. If the entropy of a random variable is interpreted as a measure of the amount of information carried by the variable, a nonzero value of the mutual information indicates that the amount of information carried by the joint random process is less than the addition of information carried by each random variable by itself. Or in other words, that the random variables had some common information in such a way that when measured as a joint process, the total amount of information is less that the addition of the information of each one. In fact, this measure has been used in several approaches of ICA as measure of the independence of the sources [13]. This is because if the obtained signals share no information (the mutual information is zero), the sources must be independent.

Applying this concept for the case of a mixture of signals, if the mutual information of the signals in a frequency bin is small, it will be indicative that there is little information sharing among the random variables involved. But if there is little information sharing is equivalent to express that the degree of mixing is small. In this way, mutual information can be used as an index of separability for the pair of signals in each frequency bin. The central bin selection will be done according to the bin that shows the lowest mutual information.

At this point we use the following assumption as in [21, 22]: For a complex valued random variable $X$, $p(x)$ is independent of the phase angle, or in other words, $p(x) = p(|x|)$. This assumption is plausible for the time evolution of a specific frequency bin, given that the STFT was calculated using arbitrary shifted windows, and the arbitrary shift affects the phase information but should not affect the pdf. In this way the mutual information between the magnitude of the signals in each bin can be estimated. To produce an estimation of the mutual information a non-parametric histogram based estimator was used [15].

There are also two other aspects to consider. On is the variation of signal levels among different bins. To make the measurement independent of these variations, we normalize the mutual information by the average magnitude of the signals of each bin. The other aspect is the effect of frequency in the parameter estimation. The parameters to estimate, particularly the delays, are obtained from the angle of the crossing terms in the mixing matrix, divided by the frequency of the bin. In this way, for the same level of accuracy in the angle estimation, a bin at higher frequencies will produce a better estimation. If the angle estimation has an error of $\zeta$, the delays have an error proportional to $\zeta/k$ where $k$ is the bin index. This means that a higher frequency bin will have less effect of the noise in the parameter estimation, thus we divide the mutual information by the frequency bin index $k$, producing lower values for higher frequencies. In this way, the optimal bin is selected as the one that minimizes the following quantity

$$J(k) = \frac{I(|z_1(\omega_k,\tau)|, |z_2(\omega_k,\tau)|}{\frac{k}{T}\sum_{i=1}^{2}\sum_{\tau=1}^{T}|z_i(\omega_k,\tau)|} \tag{12}$$

where $T$ is maximum frame index used in the STFT.

### 3.2 Correlated Wiener Postfilter

The Wiener postfilter used in [8] has shown to be very usefull, but in its simple form of Eq. 7 a lot of information available in the source and noise estimation is disregarded. One of the most important effects of reverberation is to propagate the information along the time. This means that some event happening at a given time will continue to have influence in future instants. In other words, the reverberation effect increases the correlation in time.

This information is not exploited in the ICA method used in this work, because the signals are assumed to be generated by random iid process. The Wiener filter proposed in [8] also does not take into account this information as the estimation of the noise is based on the current time only. But for a batch method, there is information available on the noise characteristics from both, past and future values, thus a more sophisticated alternative can be implemented. In addition, the obtained signals after separation can have an arbitrary delay. That is, there is nothing that guarantees synchronization of the extracted sources, thus the information used as estimation of noise in the original Wiener filter could be related to a different instant than that for which was used.

These two aspects motivate us to explore some way to introduce the time correlation information in the noise estimation. To achieve this, the Wiener time frequency postfilter is modified in the following way

$$F_{\mathscr{W},1}(\omega,\tau) = \frac{|z_1(\omega,\tau)|^2}{|z_1(\omega,\tau)|^2 + \sum_{k=-p}^{p} c_k\,|z_2(\omega,\tau-k)|^2}\,, \tag{13}$$

where $k$ represents the index of lag, $p$ is the maximum lag to consider, and $c_k$ are properly chosen weights that must take into account amount of contribution of the noise in that lag to the noise present in the source. The second term in the denominator represent an estimation of the noise in the present time, given past and future information of the corresponding bin. This produces a more accurate estimation of the noise, and although it considers a noncausal estimation, it must be noted that even the basic Wiener postfilter is noncausal, and this is feasible for batch algorithms.

The important aspect here is how to fix the weighting constants $c_k$. These weights should be large if the delayed version of the noise has an important effect in the current time, otherwise it should be small. The effect of delayed versions of the noise can be evaluated by some measure of similitude with respect to the noisy signal. To calculate such a similitude we use the correlation among the accumulated squared magnitude over all frequencies. These accumulated squared magnitudes are given by

$$\varepsilon_{z_i}(\tau) = \sum_{j=1}^{L}\left|z_i(\omega_j,\tau)\right|^2 \tag{14}$$

where $j$ is the frequency bin index and $L$ the index of the maximum frequency. With this definition, the weight coefficients are defined as the normalized correlation

$$c_k = \frac{\sum_\tau \varepsilon_{z_1}(\tau)\varepsilon_{z_2}(\tau+k)}{\|\varepsilon_{z_1}\|\|\varepsilon_{z_2}\|},\ \forall -p \le k \le p\,. \tag{15}$$

with an equivalent definition for the filter to enhance the other source, interchanging the roles of $z_1$ and $z_2$.

The value of $p$ is related to two factors. One is the already mentioned reverberation. The longer the reverberation time of the room, the larger the number of successive windows that will be important in the estimation. Also, the amount of overlapping between windows in the STFT increases the redundancy. In PMBSS an overlapping factor of 50% is used, and thus this aspect will have a minimal effect in the optimal value of $p$.

## 4 Results and Discussion

The performance of the proposed methods was evaluated using two different quality measures. One is the Perceptual Evaluation of Speech Quality (PESQ) measure, an objective method defined in the standard ITU P.862 for evaluation of communication channels and speech codecs. In a series of studies, this measure was found to be highly correlated with the output of speech recognition systems, when the input was preprocessed by fd-ICA methods [6, 7].

The other evaluation was performed using an automatic speech recognition system. This is a state-of-the-art continuous speech recognition system based on semi-continuous hidden Markov models, with context independent phonemes in the acoustic models, using Gaussian mixtures and bigram language model estimated from the transcriptions. The front-end was Mel Frequency Cepstral Coefficients (MFCC), including energy and the first derivative of the feature vector. The system was built using the HTK toolkit [26].

The audio material for the experiments was taken from a subset of the Spanish speech Albayzin database [16], and we also used white noise from Noisex-92 database [24]. All the material uses a sampling frequency of 8 kHz. The acoustic model was trained using 585 sentences from a subset related to Spanish geography questions. A set of 5 sentences uttered by two male and two female, for a total of 20 utterances, was used to evaluate the speech recognition rate.

The mixtures were recorded in a real room as in Fig. 4. This room has 4 × 4.9 m with a ceiling height of 2.9 m. The room has a reverberation time of $\tau_{60} = 120$ ms, but plywood reverberation boards were added in two of the room walls to increase this time to $\tau_{60} = 200$ ms. Two loudspeakers were used to replay the sound sources and the resulting sound field was captured with two measurement omnidirectional microphones spaced by 5 cm. The 20 sentences were mixed with the two kind of noises, at two different power ratios: 0 dB and 6 dB. In this way there are four sets of mixtures of the 20 test sentences.

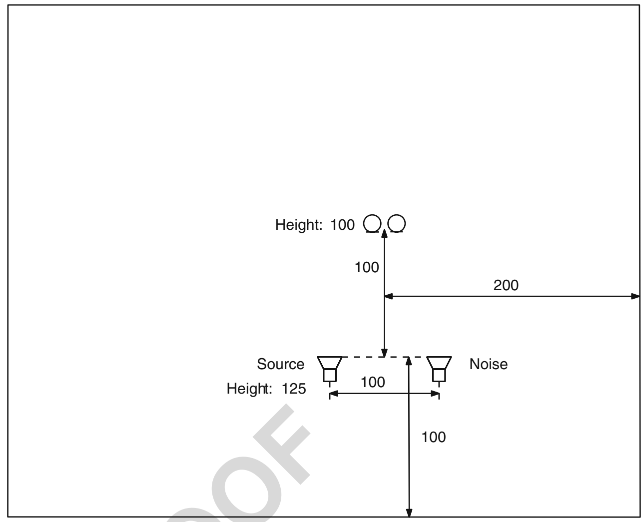


**Figure 4** Room setup used in the mixtures generation. All dimensions are in cm.



The recognition performance was evaluated using the word recognition rate, calculated after forced alignment of the system transcription with respect to the reference transcription. This measure was calculated in the standard way as

$$WRR\% = \frac{N - S - D}{N} 100\% \, , \tag{16}$$

where $N$ is the total number of words in the reference transcriptions, $S$ is the number of substitution errors, and $D$ is the number of deletion errors [26].

For the standard PMBSS we used the same configuration as proposed in the previous work, with central bin fixed at 3/8 of the maximum frequency for white noise, and 5/8 of the maximum frequency for speech noise. In all experiments we fixed the number of lateral bins to use in 10.

### 4.1 Optimal Lag for the Wiener Postfilter

The proposed Wiener postfilter depends on one parameter that needs to be determined: the maximum number of lags $p$ to consider in the noise estimation. There is a compromise in the selection of this parameter. On one side, if the reverberation time is long, the information of the noise in one instant will have importance at a wider ranges of time instants, and thus a larger $p$ should be used. On the other side, if too much lags are combined, there is an increasing probability of having time-frequency tiles for which both, the estimated source and the estimated noise, have significant energy, and this will produce a degradation on the source estimation. To verify the influence of this

Q3

**Table 1** Average separation quality as function of the number of lags used to estimate the Wiener filter.

| Power | Noise | STD | $p = 0$ | $p = 1$ | $p = 2$ | $p = 3$ |
|---|---|---|---|---|---|---|
| 6 dB | Speech | 2.74 | 2.74 | **2.80** | 2.78 | 2.73 |
| | White | 2.84 | 2.83 | **2.88** | 2.86 | 2.83 |
| 0 dB | Speech | 2.50 | 2.48 | **2.52** | 2.45 | 2.41 |
| | White | 2.59 | 2.54 | **2.67** | 2.66 | 2.65 |
| Average | | 2.67 | 2.65 | **2.71** | 2.69 | 2.65 |

parameter, the set of 20 test mixtures, under the two kind of noises and the two noise powers, were separated using values of 0, 1, 2 and 3 for $p$, and the PESQ quality evaluated on each separated source. For comparison we used also the standard method (STD) as proposed in [8]. Table 1 presents the results.

As it can be seen, the best results are obtained for a maximum lag of 1. The use of $p = 0$ imply using as noise estimation only the present time instant, which would be the same as in the standard PMBSS method. The difference is in the use of weights, that being lower than one will reduce the noise estimation with respect to the standard method where this weight is always equal to one. When the number of lags considered is increased, the quality is lowered. This is due to the increasing distortions introduced by the Wiener postfilter when it eliminates more and more frequency components. Nevertheless, it must be noted that when the sources are heard, the competing source is almost completely eliminated, but the resulting spectrogram show an increased number of gaps due to the excessive elimination of frequency components, which produce the reduction on PESQ.

This effect in the spectrogram can also be seen in Fig. 5. To generate this figure, the magnitude of the Wiener postfilter was draw in colorscale, for $p = 0, 1, 2,$ for one example of speech-speech mixture at 0 dB. Also the spectrogram of the original (desired) source is shown. The effect of adding lags is a sharpening in the spectral characteristic of the desired source. As the number of lags is increased, the Wiener filter approaches a binary mask with sharp transitions, which provides better rejection of the undesired source, but also introducing distortions in the desired source. On the contrary, for small $p$ the shape is smoother, with better preservation of the desired source, but a greater leakage of the undesired one.

## 4.2 Evaluation of the Bin Selection Method

To show that the proposed method can properly select the optimum bin, we have chosen four examples of mixtures, two with speech and the other two with white noise as competing sources, all at 0 dB of power ratios. The separation method was applied using a fixed number of 10 lateral bins at each side of the selected central bin to estimate the mixing parameters. A window length of 256 samples with window shift of 128 samples was used. This produces a transform with 129 bins. The central bin was varied from 11 to 118, and for each value of the central bin, the basic separation method was applied and the PESQ score over the whole reconstructed signal was calculated. In this way, a graphic of the achieved quality in function of the central



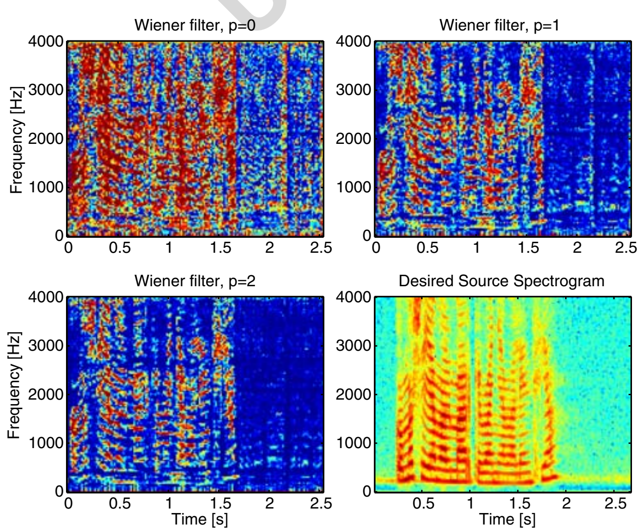


**Figure 5** Effect of the number of lags $p$ in the Wiener filter. For reference, the desired source spectrogram is also shown.

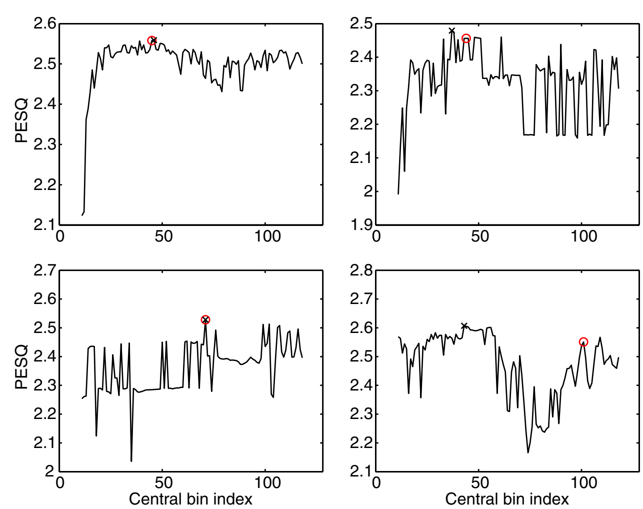


**Figure 6** Automatic central bin selection examples. The PESQ as a function of the central bin is drawn. The maximum PESQ is marked with a *cross*, and the quality of the automatic selected bin with a *circle*.

**Table 2** Average separation quality (PESQ) for the different methods evaluated in this work and the mixtures.

| Power | Noise | Mix | STD | BIN | WIENER | FULL |
|---|---|---|---|---|---|---|
| 6 dB | Speech | 2.11 | 2.74 | 2.83 | 2.80 | **2.89** |
| | White | 1.98 | 2.84 | 2.83 | **2.88** | 2.87 |
| 0 dB | Speech | 1.73 | 2.50 | 2.60 | 2.52 | **2.65** |
| | White | 1.64 | 2.59 | 2.56 | 2.67 | **2.63** |
| Average | | 1.86 | 2.67 | 2.70 | 2.71 | **2.76** |

bin can be done. Then, the proposed method is applied, and the automatically selected bin reported. This allows to verify if the method can identify the optimum bin properly.

Figure 6 show the results. The first row has two examples of the PESQ for the case of white noise, and the second row the same measure for the case of speech noise. In each case, a cross marks the best PESQ value possible, and a circle mark the obtained PESQ with the automatically selected bin. It can be seen that usually the method is able to find the bin which produces the optimum PESQ, and when it cannot, it detects a bin that produces a local maximum in quality.

### 4.3 Comparative Evaluation

Finally we present the results of PESQ score and word recognition rate for the different alternatives of the method: the standard PMBSS method (STD), the method with only the central bin selection changed (BIN), the method with central bin fixed but with the improved Wiener postfilter (WIENER), and the full proposed method (FULL). Tables 2 and 3 present the results for PESQ and WRR% respectively, for the evaluated methods and also for the mixtures without any processing (that is, as they are captured by the microphones).

The results show that both proposed methods provide for an improvement in the quality of the separated signals, which is reflected in both, improvements in PESQ and in WRR. Moreover, when the two methods are applied together the improvement is even larger than the improvements obtained by the separated methods. This is clearly seen the PESQ average results, where the individual improvements are of 0.03 and 0.04, but combined contribute to a global 0.09 improvement. The complete method provides for a 6% relative improvement in quality measured as PESQ score, and an increase of 1.64% in the average recognition rate. It must be noted that the processing time is almost not changed by these new alternatives (only about 5% increase in processing time), and thus the method maintains its very high processing speed.

**Table 3** Word recognition rates (WRR%) for the different methods evaluated in this work and the mixtures.

| Power | Noise | Mix | STD | BIN | WIENER | FULL |
|---|---|---|---|---|---|---|
| 6 dB | Speech | 44.50 | 84.66 | **86.00** | 84.13 | 85.19 |
| | White | 19.54 | 84.00 | **84.50** | 82.50 | 80.50 |
| 0 dB | Speech | 30.00 | 82.50 | 83.00 | 84.66 | **86.00** |
| | White | 7.20 | 67.50 | 70.00 | **73.50** | **73.50** |
| Average | | 25.31 | 79.66 | 80.87 | 81.20 | **81.30** |

## 5 Conclusions

In this work, the PMBSS method was analyzed with increased detail, providing insights in the reason why it is very successful in achieving separation and some reverberation reduction. In particular it was shown why this reverberation reduction is produced even when the separation model is supposed to produce separation but not reverberation reduction.

This paper also addresses an aspect that was left for future work in [8], which is the selection of the optimal central bin to be used in the estimation of the mixing parameters stage. This selection is automatically done by means of an estimation of mutual information, which is used as a measure of the amount of mixing in each bin, using then the bin which shows less mixed signals.

Finally the Wiener postfilter was improved, taking into account the temporal correlation introduced by the reverberation. The noise estimation was done by a weighted average of lagged spectra, where the proper weights are selected by a cross correlation.

The proposed methods were evaluated by means of an objective quality measure and a speech recognition system. The method for central bin selection is capable of detecting the optimal central bin. The two proposed methods produced better objective quality of the obtained signals, and improvements in the recognition rate.

## References

1. Araki, S., Mukai, R., Makino, S., Nishikawa, T., & Saruwatari, H. (2003). The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing, 11*(2), 109–116.
2. Benesty, J., Makino, S., & Chen, J. (Eds.) (2005). *Speech enhancement. Signals and communication technology*. New York: Springer.
3. Brandstein, M., & Ward, D. (Eds.) (2001). *Microphone arrays: Signal processing techniques and applications* (1st ed.). New York: Springer.

4. Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing. Learning algorithms and applications*. New York: Wiley.
5. Coverm, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). New York: Wiley-Interscience.
6. Di Persia, L., Yanagida, M., Rufiner, H. L., & Milone, D. (2007). Objective quality evaluation in blind source separation for speech recognition in a real room. *Signal Processing, 87*(8), 1951–1965.
7. Di Persia, L., Milone, D., Rufiner, H. L., & Yanagida, M. (2008). Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing, 88*(10), 2578–2583.
8. Di Persia, L. E., Milone, D. H., & Yanagida, M. (2009) Indeterminacy free frequency-domain blind separation of reverberant audio sources. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(2), 299–311.
9. Douglas, S. C., & Sun, X. (2003). Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication, 39*(1–2), 65–78.
10. Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication, 16*(3), 261–291.
11. Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation, 17*(9), 1875–1902.
12. Huang, Y., Benesty, J., & Chen, J. (2006). *Acoustic MIMO signal processing* (1st ed.). New York: Springer.
13. Hyvärinen, A., & Oja, E. (2000) Independent component analysis: Algorithms and applications. *Neural Networks, 13*(4–5), 411–430.
14. Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication, 19*(22), 1–15.
15. Moddemeijer, R. (1999). A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations. *Signal Processing, 75*(1), 51–63.
16. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., et al. (1993). *Albayzin speech database design of the phonetic corpus*. Tech. rep., Universitat Politècnica de Catalunya (UPC), Dpto. DTSC.
17. Murata, N., Ikeda, S., & Ziehe, A. (2001). An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing, 41*, 1–24.
18. Omologo, M., Svaizer, P., & Matassoni, M. (1998) Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication, 25*(1–3), 75–95.
19. Parra, L., & Spence, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing, 8*(3), 320–327.
20. Rufiner, H. L., Torres, M. E., Gamero, L., & Milone, D. H. (2004). Introducing complexity measures in nonlinear physiological signals: Application to robust speech recognition. *Physica A: Statistical Mechanics and its Applications, 332*(1), 496–508.
21. Sawada, H., Mukai, R., Araki, S., & Makino, S. (2002). Polar coordinate based nonlinear function for frequency-domain blind source separation. In *IEEE international conference on acoustics, speech, and signal processing* (Vol. 1, pp. I–1001–I–1004).
22. Sawada, H., Mukai, R., Araki, S., & Makino, S. (2003). Polar coordinate based nonlinear function for frequency-domain blind source separation. *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences, E86-A*(3), 590–596.
23. Sawada, H., Mukai, R., Arakim S., & Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing, 12*(5), 530–538.
24. Varga, A., & Steeneken, H. (1993). Assessment for automatic speech recognition II NOISEX- 92: A database and experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247–251.
25. Yilmaz, O., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing, 52*(7), 1830–1847.
26. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., et al. (2005). *The HTK book (for HTK version 3.3)*. Cambridge: Cambridge University Engineering Department.

JrnlID 11265_ArtID 443_Proof# 1 - 10/12/09

# AUTHOR QUERIES

**AUTHOR PLEASE ANSWER ALL QUERIES**

Q1. Please provide biographies and photographs of the authors.
Q2. Please check if the authors' affiliations were presented correctly.
Q3. Please provide an explanation for the significance of the data presented in bold in Tables 1–3.