Pseudoanechoic blind source separation with improved Wiener postfilter *

Leandro E. Di Persia[†] Diego H. Milone[†] Masuzo Yanagida[‡]

†*Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Santa Fe, Argentina*
{ldipersia,dmilone}@fich.unl.edu.ar
‡*Department of Intelligent Information Engineering and Science, Doshisha University, Kyotanabe, Japan*
myanagid@mail.doshisha.ac.jp

***Abstract*— The pseudoanechoic model was proposed recently to simplify the parameter estimation in blind source separation based on frequency-domain independent component analysis. In the method, after separation based in the pseudoanechoic model a time-frequency Wiener postfilter to improve the separation is applied. In this contribution, a deeper analysis of the working principles of the Wiener postfilter is presented. Furthermore, a variation of this postfilter to improve the performance using the information of previous frames is introduced. The improvements obtained through the new method 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.**

***Keywords*— Pseudoanechoic model, Blind source separation, Automatic speech recognition, Wiener postfilter**

1. INTRODUCTION

The speech recognition systems trained under laboratory conditions, suffer a strong degradation in their performance when used in real environments [9, 13]. Several aspects contribute to this degrading effect. One of them is related to the use of distant microphones that allows for other sources to arrive with equivalent powers to the microphones, and moreover, allows the arrival of echoes of the desired source [11]. 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 different levels of the speech recognition system [7]. This work is focused in the preprocessing of 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.

This work is focused in a recently proposed frequency-domain independent component analysis (fd-ICA) algorithm, which uses a pseudoanechoic 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 environments where other approaches fail, and with a very high processing speed [6]. This contribution will be focused in producing an improvement in the postfiltering stage of the PMBSS method.

In the following section, a revision of PMBSS will be presented, including a new analysis of the working principles of the Wiener postfilter. Next, an alternative method for the Wiener postfilter, exploiting the temporal information to improve the noise estimation will be introduced. This section is followed by a series of experiments to show the improvements introduced by the proposed method. Then, general conclusions are presented.

2. PSEUDOANECHOIC MODEL FOR BSS

The blind source separation task 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 [8]. 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 $j = 1, \ldots, M$, the mixture can be represented at each instant t as [3]:

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

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

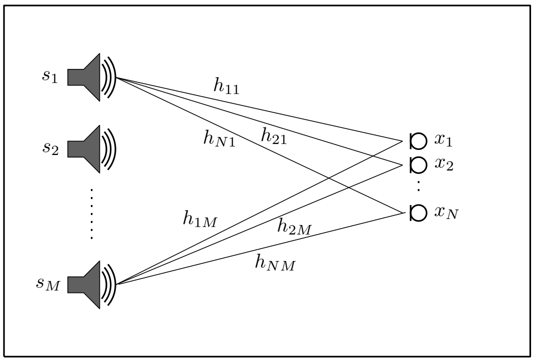

Figure 1: A case of cocktail party with M sources and N microphones.

where $*$ stands for convolution.

Let us form a vector of sources,

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

and the same for the vector of mixtures $\mathbf{x}(t)$ measured by the microphones, where $[\cdot]^T$ stands for transposition. Then Eq. (1) can be written as:

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

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.

One of the most successful approaches to solve this problem is based on the search of statistical independence of the obtained sources in the time-frequency domain. This is called the frequency-domain independent component analysis method (fd-ICA) [14]. If a short time Fourier transform (STFT) is applied to Eq. (2), the mixture can be written as [2, chapter 13]

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

where the variable τ represents the time localization given by the sliding window in the STFT, and ω 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.

2.2. The pseudoanechoic model

In a previous development [6], 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

$$\begin{aligned} x_1(t) &= s_1(t) * h_{11}(t) + s_2(t) * h_{12}(t) \\ x_2(t) &= s_1(t) * h_{21}(t) + s_2(t) * h_{22}(t) \ . \end{aligned} \tag{4}$$

Under the assumption of closely spaced microphones, the crossing impulse response can be expressed as a delayed and scaled version of the direct impulse response, 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

$$\mathbf{x}(\omega, \tau) = \begin{bmatrix} 1 & \beta e^{-jd_2\omega} \\ \alpha e^{-jd_1\omega} & 1 \end{bmatrix} \mathbf{z}(\omega, \tau) \ , \tag{5}$$

where now the $\mathbf{z}(\omega, \tau)$ contains the reverberant sources, $z_i(\omega, \tau) = h_{ii}(\omega) s_i(\omega, \tau)$. 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 (similar to the one used in 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 the pseudoanechoic mixing model, the PMBSS algorithm was introduced. Simply speaking, this method aims to produce the $\mathbf{z}$ sources mentioned before. In Eq. (5), the mixing matrix is easy to synthesize. For all frequencies, the parameters α, β, d_1 and d_2 have constant values, if they can be robustly identified for one frequency, they can be used to synthesize the mixing matrix and 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 matrices for all frequencies using the estimated parameters, and separation; 3) Application of a time-frequency Wiener postfilter.

One 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 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

$$F_{\mathcal{W},1}(\omega, \tau) = \frac{|z_1(\omega, \tau)|^2}{|z_1(\omega, \tau)|^2 + |z_2(\omega, \tau)|^2} \ , \tag{6}$$

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 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 a pair of null beamformers, where each beamformer reject the signals arriving from the estimated direction of arrival of each of the sources.

In an environment with no reverberation, if one of the sources is eliminated, the resulting signal will have information only of the other source, and a good separation will be obtained. But in reverberant environments, there are echoes arriving to the array from other directions that 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.

A 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 ϕ, given by

$$\mathbf{v}(f,\phi) = \left[1, e^{\frac{-j2\pi f d \sin(\phi)}{c}}, \cdots, e^{\frac{-j2\pi f (N-1) d \sin(\phi)}{c}}\right]^T, \tag{7}$$

where d is the microphone spacing and c the sound speed. This array response vector characterices 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

$$r(f,\phi) = \mathbf{a}^H \mathbf{v}(f,\phi)\,, \tag{8}$$

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.

Let us assume that there is a sound field produced by white and stationary signals, with equal power from all directions. In this case, the behaviour of the combined separation and Wiener filter process can be analyzed using the beampatterns, as the beampattern

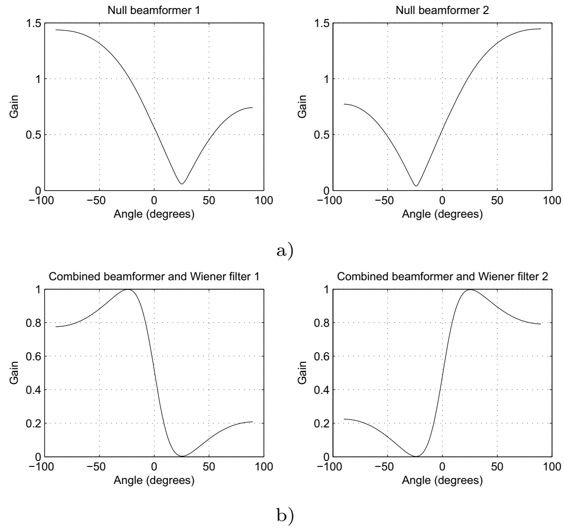

Figure 2: Effect of the Wiener postfilter on the beampatterns. a) the beampatterns generated from the separation matrix. b) the beampatterns after application of the Wiener filter.

output will be the actual magnitude at the output of the separation, as a function of the arrival angle.

Figure 2 shows the beampatterns obtained from the separation matrix in the bin corresponding to 2000 Hz (for other frequencies the analysis is equivalent), for two sources located at ± 26 degrees. Part a) 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, and in the direction of the other source the gain tends to zero. In part b), 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.

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 that the Wiener postfilter helps in improving the fundamental limitation of the fd-ICA approach as analyzed in [1], that is, the impossibility to reject the echoes.

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. 2 represents the worst case of possible inputs, and thus for more realistic cases an even better behaviour can be expected.

3. CORRELATED WIENER POSTFILTER

The proposed Wiener postfilter has shown to be very usefull, but in its simple form of Eq. (6) 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 originally proposed Wiener filter also does not take into account this information as the estimation of the noise is based on the current time only. In addition, 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

$$F_{\mathcal{W},1}(\omega,\tau) = \frac{|z_1(\omega,\tau)|^2}{|z_1(\omega,\tau)|^2 + \sum_{k=-p}^{p} c_k\, |z_2(\omega,\tau-k)|^2}\,, \tag{9}$$

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 the amount of contribution of the noise in that lag, to the noise present in the source.

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

$$\epsilon_{z_i}(\tau) = \sum_{j=1}^{L} |z_i(\omega_j,\tau)|^2 \tag{10}$$

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

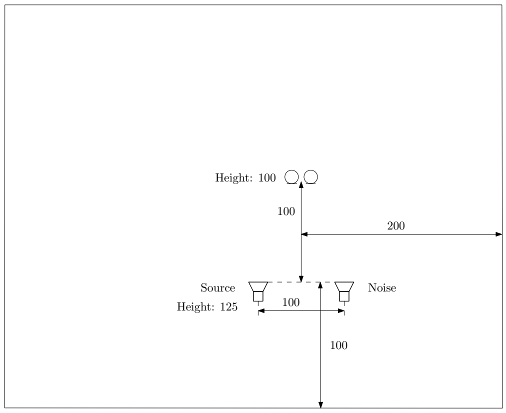

Figure 3: Room setup used in the mixtures generation. All dimensions are in cm.

$$c_k = \frac{\sum_\tau \epsilon_{z_1}(\tau)\epsilon_{z_2}(\tau+k)}{\|\epsilon_{z_1}\|\|\epsilon_{z_2}\|}\,,\ \forall\ -p \le k \le p\,, \tag{11}$$

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

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 that was found to be highly correlated with the output of speech recognition systems, when the input was preprocessed by fd-ICA methods [5, 4].

The other evaluation was performed using an ASR system. This is a 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 [16].

The audio material for the experiments was taken from a subset of the Spanish speech Albayzin database [10], and we also used white noise from Noisex-92 database [15]. 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. 3. This room has 4 x 4.9 m with a ceiling height of 2.9 m. The room has a reverberation time of $\tau_{60} = 200$ milliseconds. 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

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

Power	Noise	PMBSS	$p = 0$	$p = 1$	$p = 2$
6 dB	Speech	2.74	2.74	**2.80**	2.78
	White	2.84	2.83	**2.88**	2.86
0 dB	Speech	2.50	2.48	**2.52**	2.45
	White	2.59	2.54	**2.67**	2.66
Ave.		2.67	2.65	**2.71**	2.69

power ratios: 0 dB and 6 dB. In this way there are four sets of mixtures of the 20 test sentences.

The recognition performance was evaluated using the word recognition rate (WRR), calculated after forced alignment of the system transcription with respect to the reference transcription. 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.

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 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 and 2 for p, and the PESQ quality evaluated on each separated source. For comparison we used also the standard method (PMBSS) as proposed in [6]. 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.

This effect in the spectrogram can also be seen in Fig. 4. 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 introduces 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.

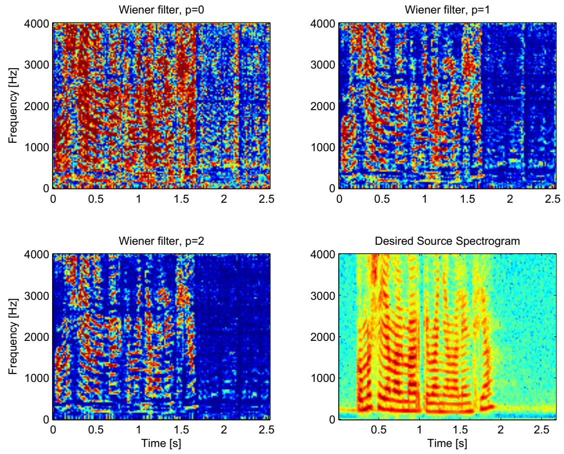

Figure 4: Effect of the number of lags p in the Wiener filter. For reference, the desired source Spectrogram is also shown.

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

Power	Noise	Mix	PMBSS	Parra	Prop.
6 dB	Speech	2.11	2.74	2.22	**2.80**
	White	1.98	2.84	2.37	**2.88**
0 dB	Speech	1.73	2.50	2.19	**2.52**
	White	1.64	2.59	2.16	**2.67**
Ave.		1.86	2.67	2.23	**2.71**

To be able of comparing the obtained results we present the results of PESQ score and word recognition rate for the state-of-the-art method (Parra)[12], the standard PMBSS method (PMBSS), and our proposed method with the modified Wiener postfilter (Prop). 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 the proposed method provide for an improvement in the quality of the separated signals, which is reflected in both, improvements in PESQ and in WRR. 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

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

Power	Noise	Mix	PMBSS	Parra	Prop
6 dB	Speech	44.50	**84.66**	49.50	84.13
	White	19.54	**84.00**	27.50	82.50
0 dB	Speech	30.00	82.50	49.00	**84.66**
	White	7.20	67.50	20.00	**73.50**
Ave.		25.31	79.66	36.50	**81.20**

method mantains its very high processing speed.

5. CONCLUSIONS

In this work, the PMBSS method was analyzed with increased detail, providing insights in the reason why it is very successfull 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.

Also, 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 correlation.

The proposed method was evaluated by means of an objective quality measure and a speech recognition system, producing better objective quality of the obtained signals, and improvements in the recognition rate.

References

[1] Shoko Araki, Ryo Mukai, Shoji Makino, Tsuyoki Nishikawa, and Hiroshi Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Trans. on Speech and Audio Proc.*, 11(2):109–116, March 2003.

[2] Jacob Benesty, Shoji Makino, and Jingdong Chen, editors. *Speech Enhancement.* Signals and Communication Technology. Springer, 2005.

[3] Andrzej Cichocki and Shun-Ichi Amari. *Adaptive Blind Signal and Image Processing. Learning Algorithms and applications.* John Wiley & Sons, 2002.

[4] Leandro Di Persia, Diego Milone, Hugo Leonardo Rufiner, and Masuzo Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Proc.*, 88(10):2578–2583, October 2008.

[5] Leandro Di Persia, Masuzo Yanagida, Hugo Leonardo Rufiner, and Diego Milone. Objective quality evaluation in blind source separation for speech recognition in a real room. *Signal Proc.*, 87(8):1951–1965, 2007.

[6] Leandro Ezequiel Di Persia, Diego Humberto Milone, and Masuzo Yanagida. Indeterminacy free Frequency-Domain blind separation of reverberant audio sources. *IEEE Trans. on Audio, Speech, and Language Proc.*, 17(2):299–311, 2009.

[7] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Comm.*, 16(3):261–291, April 1995.

[8] Yiteng Huang, Jacob Benesty, and Jingdong Chen. *Acoustic MIMO Signal Processing.* Springer, 1 edition, October 2006.

[9] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Comm.*, 19(22):1–15, 1997.

[10] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, and C. Nadeu. Albayzin speech database design of the phonetic corpus. Technical report, Universitat Politècnica de Catalunya (UPC), Dpto. DTSC, 1993.

[11] M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Comm.*, 25(1-3):75–95, August 1998.

[12] Lucas Parra and Clay Spence. Convolutive blind separation of non-stationary sources. *IEEE Trans. on Speech and Audio Proc.*, 8(3):320–327, 2000.

[13] Hugo Leonardo. Rufiner, María Eugenia. Torres, Lucas Gamero, and Diego Humberto. Milone. Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition. *Physica A: Stat. Mechanics and its Applications*, 332(1):496–508, February 2004.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. on Speech and Audio Proc.*, 12(5):530–538, 2004.

[15] A. Varga and H. Steeneken. Assessment for automatic speech recognition II NOISEX- 92: A database and experiment to study the effect of additive noise on speech recognition systems. *Speech Comm.*, 12(3):247–251, 1993.

[16] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK book (for HTK Version 3.3).* Cambridge University Engineering Department, Cambridge, 2005.