SPEECH EXTRACTION BASED UPON A COMBINED SUBBAND INDEPENDENT COMPONENT ANALYSIS AND NEURAL MEMORY

Tetsuya Hoya [1], *Allan Kardec Barros* [2], *Tomasz Rutkowski* [1], *and Andrzej Cichocki* [1]

[1]Laboratory for Advanced Brain Signal Processing,
BSI RIKEN, 2-1, Hirosawa, Wakoh-City, Saitama 351-0198, Japan
[2]Department of Electrical Engineering, University of Maranhão (UFMA), Brasil
e-mail: hoya@bsp.brain.riken.go.jp

ABSTRACT

This paper presents a novel approach for speech extraction by a combined subband independent component analysis and neural memory. In the approach, probabilistic neural networks followed by the subband independent component analysis processing units are used for the neural memory to identify firstly the speaker and then compensate for the 'side-effects', i.e., the scaling and the permutation disorder, both of which are particularly problematic for subband blind extraction. Simulation study shows that the combined scheme can effectively extract the speech signal of interest from the instantaneous / delayed mixtures, in comparison with the conventional subband/fullband approaches.

1. INTRODUCTION

The *cocktail party problem* can be generally solved; i.e., in a place where a group of people are talking simultaneously, we can still be selectively attentive to a particular person and communicate with each other. However, it is well-known that to imitate such capability inherent to humans is a challenging topic. In the last decade, with the advancements in the algorithms for independent component analysis (ICA) [1], emulation of this facility by a machine has been tackled by a number of researchers [2].

In recent studies [3] - [5], a variant of subband blind extraction approaches was proposed. As reported, these methods work well to extract the highest energy speech component but the enhanced speech, so obtained, is greatly deteriorated. This is mainly due to the incomplete reconstruction of the enhanced signal, resulting from both the scale misadjustment and permutation indeterminacy in the separated subband signals.

In this paper, we therefore propose a combined scheme for subband blind extraction with a neural memory consisting of a multiple of probabilistic neural networks (PNNs) [6], so as to compensate for the incomplete reconstruction. In contrast, in [7], a neural approach for the cocktail party problem is proposed. The approach exploits so-called a 'cortronic' neural network (i.e., correlation matrix trained by a Hebbian-type learning algorithm) for the extraction of one speaker. The manner in which neural memory is utilised in this paper is rather different in that the neural network itself extracts the speech signal. Moreover, the combined scheme benefits from the flexible configuration property of PNNs [8].

2. METHOD

In cocktail party situations, it is assumed that there are M source speech signals at time k, i.e., $s_i(k)$, $i = 1, 2, \cdots, M$ which can be represented in vector form as $\boldsymbol{s}(k) = [s_1(k), s_2(k), \cdots, s_M(k)]^T$, where $[\cdot]^T$ denotes the vector transpose operator, and the two signals arriving at the two sensors (microphones) at discrete time k, $\boldsymbol{u}_i(k) = [u_1(k), u_2(k)]^T$, can be defined as an under-determined linear convolutive model:

$$\boldsymbol{u}_i(k) = \sum_{n=-\infty}^{\infty} \boldsymbol{H}(n)\boldsymbol{s}(k-n) \qquad (1)$$

where $\boldsymbol{H} \in \Re^{2\times M}$ is a linear filter operator and defines the mixture. In the real acoustic environment,

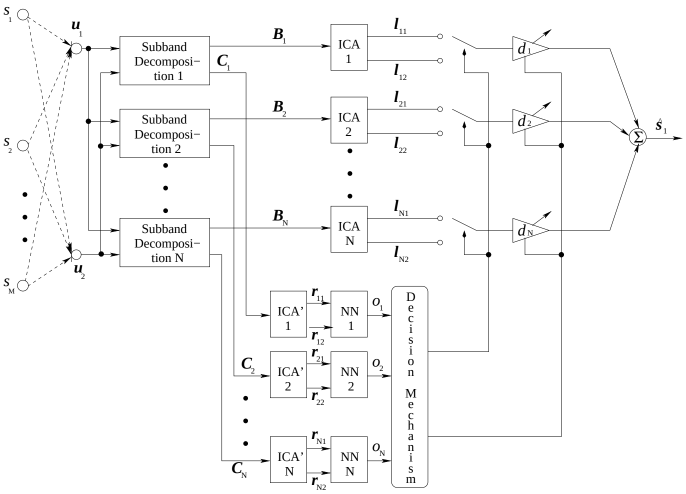

Fig. 1. Block diagram of the combined subband blind speech extractor containing neural memory units denoted by NN$_1$ to NN$_N$.

$\boldsymbol{H}$ is generally a non-minimum phase low-pass filter [9] and recovering the original signals is thus very hard.

Fig. 1 illustrates the block diagram of the proposed speech extraction scheme. Note that, as implicitly denoted by the variables in Fig. 1, the proposed method works in a batch operated manner rather than on-line. (Thus, in practice, the signals are given in vector forms, e.g., $\boldsymbol{u}_i(k) = [u_i(k), u_i(k-1), \cdots, u_1(k-L+1)]^T$, representing a length L window.) In the figure, note also that the neural memory unit for each subband NN i ($i = 1, 2, \cdots, N$) is newly introduced and that, after the identification by the neural memory units, the decision mechanism 1) determines which separated subband signal $\boldsymbol{l}_{ij}$ obtained by ICA i should be used for the summation operation and 2) outputs the scaling adjustment factors d_i to finally obtain the reconstructed speech $\hat{\boldsymbol{s}}_1$ as

$$\hat{\boldsymbol{s}}_1 = \sum_{i=1}^{N} d_i \boldsymbol{l}_{ip} \tag{2}$$

where $p = 1$ or 2 and is determined by the reference to the neural memory.

2.1. The Subband ICA Mechanism

As stated in [5], the effectiveness of subband blind extraction resides in the property that narrow band signals are less prone to the convolutive effects in comparison with the original fullband signal. In this paper, the subband coding mechanism developed in [5] is considered. The subband coding scheme [5] is based upon the concept of harmonicity of the voiced sounds, which has also been exploited in some models of *computational auditory scene analysis* (CASA) [10]. In summary, the mechanism involves two steps: 1) extraction of the fundamental frequencies obtained by applying both the wavelet and Hilbert transforms and 2) the bank of adaptive band-pass filters centered at the fundamental frequency $f0$ obtained in 1) and its harmonics.

In the proposed scheme, both the two-channel subband signals $\boldsymbol{B}_i$ ($\Re^{L\times 2}, i = 1, 2, \cdots, N$, in block form) and instantaneous amplitude envelope signals $\boldsymbol{C}_i$ ($\Re^{L'\times 2}$, $i = 1, 2, \cdots, N$), which can be obtained as intermediary signals in 2) above, are exploited. The reason for using $\boldsymbol{C}_i$ instead of $\boldsymbol{B}_i$ for the reference to the neural memory is based on the assumption that the amplitude envelope retains sufficient information about the original speech, whilst its column vector length, L', is less than that of each column vector in $\boldsymbol{B}_i$ [5] (which is proportion to 1/N, N: the number of subbands), which is desirable for quick data processing by the PNNs.

For the blind extraction of each subband signal, we apply the second-order blind identification (SOBI) algorithm [11] which is based upon joint diagonalisation of correlation matrices and known to be robust for nonstationary signals such as speech. As in Fig. 1, the SOBI algorithm is independently applied to both the separation of the subband signal $\boldsymbol{B}_i$ (ICA i) and the amplitude envelope $\boldsymbol{C}_i$ (ICA' i). (In a noisy environment, the robust form of SOBI [1] can also be exploited.)

Then, it is considered that, due to the statistical invariance between the subband signal $\boldsymbol{B}_i$ and the amplitude envelope $\boldsymbol{C}_i$, there is no ordering problem between the corresponding separated signals $\boldsymbol{l}_{ij}$ and $\boldsymbol{r}_{ij}$ ($j = 1, 2$), i.e., both $\boldsymbol{l}_{ij}$ and $\boldsymbol{r}_{ij}$ correspond to the same target source signal. Although the theoretical justification is still under investigation, this was empirically confirmed by the preliminary simulation study.

2.2. The Probabilistic Neural Network

In Fig. 1, each unit 'NN i' represents a distinct probabilistic neural network (PNN). The PNN [6] is a family of radial basis function neural networks (RBF-NNs) [12] and reformulation of kernel discriminant analysis [13] in the artificial neural network context.

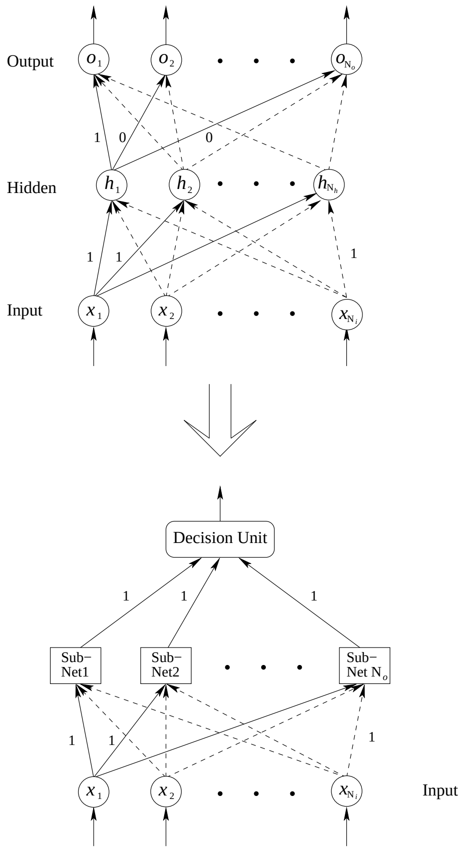

Fig. 2. Illustration of the topological equivalence between a conventional PNN (upper) and that realised modular form (lower).

Recently, the utility of PNN/ generalised regression neural networks (GRNNs) has been increased especially in pattern classification, due to its straightforward and flexible configuration (i.e., network growing/shrinking) property (e.g., see [14]) and robustness, in comparison with the commonly used multilayered perceptron neural networks [12] trained by a backpropagation type algorithm [15]. Moreover, in [8] it is reported that a PNN even exhibits a capability in accommodating new classes.

In Fig. 2, each input neuron x_i ($i = 1, 2, \cdots, N_i$) corresponds to the element in the input vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_{N_i}]^T$, h_j ($j = 1, 2, \cdots, N_h$) is the j-th RBF (note that N_h is variable), $\| \cdots \|_2^2$ denotes the squared L_2 norm, and the output neuron o_k ($k = 1, 2, \cdots, N_o$) is given as

$$o_k = \frac{1}{\delta} \sum_{j=1}^{N_h} w_{j,k} h_j, \qquad (3)$$

where $\delta = \sum_{k=1}^{N_o} \sum_{j=1}^{N_h} w_{j,k} h_j$, $\boldsymbol{w}_j = [w_{j,1}, w_{j,2}, ..., w_{j,N_o}]^T$, and

$$h_j = exp(-\frac{\|\boldsymbol{x} - \boldsymbol{y}_j\|_2^2}{\sigma_j^2}), \qquad (4)$$

where $\boldsymbol{y}_j$ is called the centroid vector, σ_j is the radius, and $\boldsymbol{w}_j$ denotes the weight vector between the j-th RBF and the output neurons. As in the upper part in Fig. 2, the structure of a PNN is similar to the well-known multilayered perceptron neural network (MLP-NN) except that RBFs are used in the hidden layer and linear functions in the output layer. In Fig. 2, when the target vector $\boldsymbol{t}(\boldsymbol{x})$ corresponding to the input pattern vector $\boldsymbol{x}$ is given as

$$\begin{aligned} \boldsymbol{t}(\boldsymbol{x}) &= (\delta_1, \delta_2, ..., \delta_{N_o}), \\ \delta_j &= \begin{cases} 1 & \text{if } \boldsymbol{x} \text{ belongs to the class} \\ & \text{corresponding to } o_k \\ 0 & \text{otherwise} \end{cases} \end{aligned} \qquad (5)$$

which assigns the weight vector between the j-th RBF and the output neurons, i.e., $\boldsymbol{w}_j = \boldsymbol{t}(\boldsymbol{x})$, the entire network eventually becomes topologically equivalent to the one with a decision unit (followed by the 'winner-takes-all' strategy) and N_o number of sub-nets as in the lower part of the figure [14]. Then, each SubNet i represents the pattern space of Class i spanned by the RBFs. In summary, the network configuration by means of a PNN is simply done as follows:

Network Growing: Set $\boldsymbol{y}_j = \boldsymbol{x}$ and fix σ_j, then add the term $w_{jk} h_j$ in (3). The target vector $\boldsymbol{t}(\boldsymbol{x})$ is used as a class 'label' indicating the subnetwork number to which the RBF belongs.

Network Shrinking: Delete the term, $w_{jk} h_j$, from (3).

As in the above, it is considered that, in hardware implementation, the network growing (learning) can be straightforwardly performed. It is well-known that the generalisation performance of RBF-NN families such as PNNs is robust, while conventional neural networks such as multilayered perceptron neural networks (MLP-NNs) with the backpropagation algorithm [15] require iterative and (quite often) long training whenever the network configuration is changed and there is always a danger of being stuck in local minima [12].

2.3. The Neural Memory

As mentioned earlier, the role of the neural memory is to determine 1) which subband signal $\boldsymbol{l}_{ij}$ should be chosen and 2) the scale adjustment factor d_i for the reconstruction.

For 1), the determination is performed as pattern recognition (identification) of each blindly extracted amplitude envelope $\boldsymbol{r}_{ij}$ (j=1,2) given to the neural memory. Thus, the input vector of the PNN is obtained from the envelope signal. In contrast, during the construction (or network growing of the PNNs) phase of the neural memory, a total of N new RBF units will be created at a time within the respective PNNs, with each $\boldsymbol{C}_i$ stored as the centroid vector (after the vectorisation) of the corresponding RBF. Then, the class ID represents the ID of the target speech signal. During the construction of the neural memory, we also store the values obtained by calculating the standard deviation of $\boldsymbol{B}_i$ as d_i in an auxiliary array.

For the reference/construction of the neural memory, there is, however, one consideration: since the data length of the amplitude $\boldsymbol{r}_{ij}$ is varied in a sample-by-sample manner, the input vector to the PNN must be normalised in not only amplitude but time (length). To adjust the length of the input vector, we considered applying a simple re-sampling mechanism with both anti-aliasing low-pass filtering and zero-padding where necessary (e.g., see [16]).

After the normalisation, pattern recognition of the input vectors $\boldsymbol{r}_{ij}$ is performed one by one by the PNN for each subband. Then, for each $\boldsymbol{r}_{ij}$, we obtain a total of $N(=$ number of subbands) recognition results (e.g., for the i-th subband, Channel j is recognised as the subband signal of the target signal $\boldsymbol{s}_1$). To finally determine which $\boldsymbol{l}_{ij}$ should be taken for the reconstruction of the target speech, we apply a simple scoring scheme: regard the channel j as the target signal if the number of subbands recognised as the target $\boldsymbol{s}_1$ is greater than that of the other channel.

For 2), now that we know which channel should be taken for the reconstruction of the target speech signal and the channel signal is identified by the neural memory, we simply recall a set of scale adjustment values d_i stored in the auxiliary array, corresponding to the identified speech. Finally, we obtain the reconstructed speech by (2).

3. SIMULATION STUDY

In the simulation study, we applied the proposed scheme and compared the performance with that of 1) the original subband approach [5] with the SOBI algorithm [11] and 2) the fullband SOBI approach. To measure objectively the extraction performance, the *energy in difference* E_{diff} was considered:

$$E_{diff} = \frac{1}{q}\sum_{i=1}^{q}\left(\frac{\hat{\boldsymbol{s}}_1}{\mathrm{std}(\hat{\boldsymbol{s}}_1)} - \frac{\boldsymbol{s}_i}{\mathrm{std}(\boldsymbol{s}_i)}\right)^2 \tag{6}$$

where q is the number of simultaneous speech signals in the mixture (known *a priori*), std$(\cdot)$ denotes the standard deviation, $\hat{\boldsymbol{s}}_1$ is the extracted signal corresponding to the target speech, and s_i are the speech signals. (Therefore, the quality of the extraction performance improves as E_{diff} approaches zero.)

For modeling the environment, we considered two types of the two-channel mixture; 1) the instantaneous and 2) the delayed mixture. For Case 2), we assumed the situation where there is one dominant speech signal with no delay and other background signals with delays and less amplitudes than the dominant speech.

3.1. Parameter Settings

In the simulation, the number $N = 64$ was chosen for the two subband approaches and the performance was tested up to seven simultaneous voices for both the instantaneous and delayed mixture cases.

For the speech signals, we collected a total of 3×4 Portuguese and 4×1 Polish speech utterances recorded at 8kHz sampling. The Portuguese speech utterances were the three digits /NOVE/, /OITO/, and /QUATRO/, each pronounced four times by two native female and a male speaker. In contrast, the four Polish utterances were the Polish words/phrases /JEDEN/, /ANONIM/, /NAZYWAM SIE/, and /KOWALSKI/, each uttered by two native female and one male speakers. Then, in the simulation, the task was to extract the Portuguese female pronounced digit /NOVE/.

For the neural memory, only three out of four utterances for each Portuguese digit (i.e., a total of nine utterances) were used for constructing the PNNs and the remaining one was used for generating the mixtures (i.e., for the reference mode). This is to simulate the memory for a single language (Portuguese). Then, the length of both the input and the centroid vector was

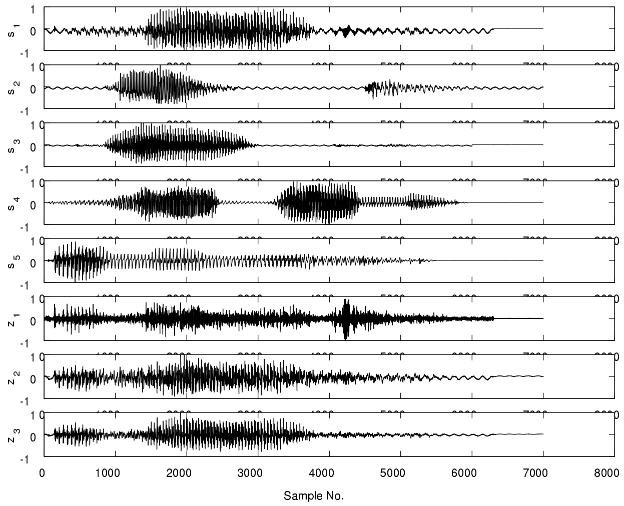

Fig. 3. Simulation Results - the Instantaneous Mixture Case with Five Simultaneous Voices (s_1 - s_5); z_1: Original Subband SOBI, z_2: Fullband SOBI, and z_3: Proposed Method.

fixed to 64 after the normalisation. For the scale adjustment values, as for each digit, three different patterns were used for the construction of PNNs, we stored the averaged values obtained from the three patterns as d_i.

3.2. Simulation Results

Due to the limit on space, we show only two sets of waveforms obtained for the instantaneous and delayed mixture case, respectively. For the instantaneous mixture, five simultaneous voices were modeled (three Portuguese $\boldsymbol{s}_1$ - $\boldsymbol{s}_3$ and two Polish $\boldsymbol{s}_4$ and $\boldsymbol{s}_5$):

$$\begin{aligned} \boldsymbol{u}_1(k) &= 0.8\boldsymbol{s}_1(k) + 0.4\boldsymbol{s}_2(k) + 0.2\boldsymbol{s}_3(k) \\ &\quad +0.1\boldsymbol{s}_4(k) + 0.3\boldsymbol{s}_5(k) \\ \boldsymbol{u}_2(k) &= 0.6\boldsymbol{s}_1(k) + 0.8\boldsymbol{s}_2(k) + 0.3\boldsymbol{s}_3(k) \\ &\quad +0.2\boldsymbol{s}_4(k) + 0.1\boldsymbol{s}_5(k) \end{aligned}$$

In the above, it was assumed that two Portuguese speakers (i.e., $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$) were dominant and the other three were background. In the simulation, Channel 1 was recognised as the voice corresponding to the Portuguese digit /NOVE/ and the speech was reconstructed according to this recognition result.

In contrast, the delayed mixture consists of three simultaneous voices (three Portuguese $\boldsymbol{s}_1$ - $\boldsymbol{s}_3$):

$$\begin{aligned} \boldsymbol{u}_1(k) &= 0.8\boldsymbol{s}_1(k) + 0.2\boldsymbol{s}_2(k-120) + 0.3\boldsymbol{s}_3(k-180) \\ \boldsymbol{u}_2(k) &= 0.6\boldsymbol{s}_1(k) + 0.2\boldsymbol{s}_2(k-125) + 0.3\boldsymbol{s}_3(k-190) \end{aligned}$$

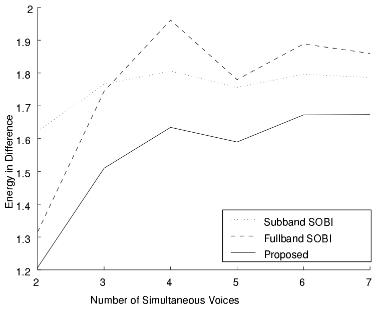

Fig. 4. Comparison of E_{diff} - the Instantaneous Mixture Case.

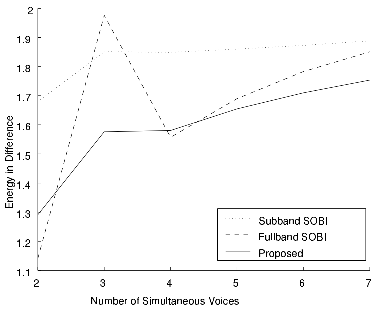

Fig. 5. Comparison of E_{diff} - the Delayed Mixture Case.

In the delayed mixture case, Channel 1 was again recognised as the target speech signal /NOVE/. As in Figs. 3 and 6, it was observed that the separation performance by the proposed method is almost comparable to the fullband SOBI and better than the original subband SOBI, and it was confirmed by the listening tests that the proposed method is the best among the three methods in terms of the quality of the extracted speech for both the instantaneous and delayed mixture cases. In Fig. 3, the effectiveness of the proposed method is noticeable between sample nos 1000 and 4000.

In Figs. 4 and 5, the comparisons of E_{diff} (defined in (6)) for the instantaneous and delayed mixture cases with varying number of simultaneous voices are respectively shown. In both the cases, it is clearly observed that the performance with the proposed method (z_3) is almost consistently superior to the other two. This is particularly remarkable for the instantaneous mixture case and was also confirmed by the informal

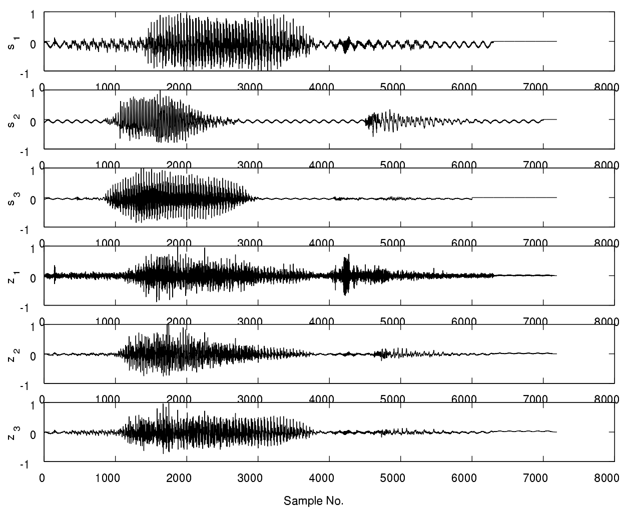

Fig. 6. Simulation Results - the Delayed Mixture Case with Three Simultaneous Voices (s_1 - s_3); z_1: Original Subband SOBI, z_2: Fullband SOBI, and z_3: Proposed Method.

listening tests.

4. CONCLUSION

In this paper, a novel subband blind speech extraction scheme with neural memory concept has been proposed. The neural memory is synthesised with PNNs applied to the subband speech envelope information. From both objective and subjective evaluation of the simulation results, it appears that the neural memory can compensate for the drawbacks within the original subband approaches and the extraction performance is consistently superior to that of other two conventional approaches. Future work includes a thorough investigation of the proposed scheme applied to the general case where convolutive mixtures are considered and is directed towards the extension of the proposed scheme to a more practical situation: continuous speech cases.

5. REFERENCES

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*, John Wiley & Sons, 2002.

[2] S. Haykin, *Unsupervised Adaptive Filtering*, John Wiley & Sons, Volume I & II, 2000.

[3] A. K. Barros, H. Kawahara, A. Cichocki, S. Kajita, T. Rutkowski, and N. Ohnishi, “Enhancement of a speech signal embedded in noisy environment using two microphones,” *Proc. ICA-2000*, Helsinki, Finland, pp. 423-428, June 2000.

[4] T. Rutkowski, A. Cichocki, and A. K. Barros, “Speech extraction from interferences in real environment using bank of filters and blind source separation,” *Proc. of the Workshop on Signal Processing Applications*, Brisbane, Australia, Dec. 2000.

[5] A. K. Barros, T. Rutkowski, F. Itakura and N. Ohnishi, “Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets,” *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 888-893, July 2002.

[6] D. F. Specht, “Probabilistic neural networks,” *Neural Networks*, no.3, pp. 109-118, 1990.

[7] S. Brian, C. N. Syrus, K. Rex, H. Raja, and H. N. Robert, “A biologically motivated solution to the cocktail party problem,” *Neural Computation*, vol. 13, no. 7, pp. 1575-1602, July 2001.

[8] T. Hoya, “On the capability of accommodating new classes within probabilistic neural networks,” accepted for publication in the *IEEE Trans. Neural Networks*, Dec. 2002.

[9] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, 2000.

[10] M. Weintraub, “A theory and computational model of auditory monaural sound separation,” Doctoral dissertation, Stanford Univ., 1985.

[11] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “Second-order blind separation of temporally correlated sources,” in Proc. Int. Conf. on Digital Sig. Proc., Cyprus, pp. 346-351, 1993.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, 1994.

[13] D. J. Hand, *Kernel Discriminant Analysis*, Research Studies Press, 1984.

[14] T. Hoya and J. A. Chambers, “Heuristic pattern correction scheme using adaptively trained generalized regression neural networks,” *IEEE Trans. Neural Networks*, vol.12, no.1, pp. 91-100, 2001.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, R. J. “Learning internal representations by error propagation,” in D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1, chapter 8, Cambridge, MA:MIT Press, 1986.

[16] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing - Principles, Algorithms, and Applications*, Macmillan, New York, 1992.