A Comparative Study of Two Different Neural Models for Speaker Recognition Systems

Geeta Nijhawan, M.K. Soni

Abstract-In recent years there has been a significant amount of work, both theoretical and experimental, that has established the viability of artificial neural networks (ANN's) as a useful technology for speech recognition. It has been shown that neural networks can be used to augment speech recognizers whose underlying structure is essentially that of hidden Markov models (HMM's). In this paper, we first give a brief overview of automatic speech recognition (ASR) and then describe the use of ANN's as statistical estimators. We have compared back propogation (BP) neural network and radial basis function (RBF) network's performance as applied to the speaker recognition. We have compared the two neural network results by MATLAB simulation. From the quantitative point we have proved that the RBF neural network is more efficient and accurate than BP neural network in speaker recognition, and thus more suitable for practical applications.

Keywords- Speaker recognition system, Linear Predictive Coding (LPC), Neural networks, Mel Frequency Cepstrum Coefficient (MFCC), Back Propagation (BP); Radial Basis Function (RBF).

I. INTRODUCTION

Speaker recognition is the task of determining a person's identity by his / her voice. This task is also known as voice recognition. It is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals or waves. This term is often confused with speech recognition. The goal of speaker recognition is to determine the speaker's identity irrespective of the words. While the goal of speech recognition is to determine what words are spoken irrespective of the speaker.

The area of speaker identification is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive in activities such as the telephone, financial transactions and information retrieval from speech databases, the utility of automatically identifying a speaker is based solely on vocal characteristic. This paper emphasizes on text independent speaker identification, which deals with detecting a particular speaker from a known population. Speaker recognition systems contain three main modules:

(1) Acoustic processing
(2) Features extraction or spectral analysis
(3) Recognition.

as shown in Fig. 1.These processes are explained in detail in subsequent sections.

Manuscript received on May 28, 2012.
Ms. Geeta Nijhawan, Deptt. of Electronics and Communications, FET, MRIU Faridabad
Dr. M.K. Soni, Executive Director & Dean, FET, MRIU, Faridabad.

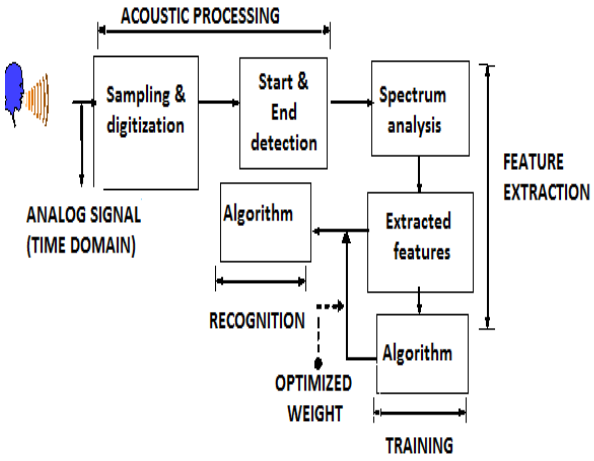

Fig.1. Basic structure of speaker recognition system

Research and development on speaker recognition methods and techniques has been undertaken for well over four decades and it continues to be an active area. Approaches have spanned from human aural and spectrogram comparisons, to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs). Researchers have applied many of techniques to speaker recognition including Hidden Markov Models (HMM) [Siohan, 1998], Gaussian Mixture Modeling (GMM) [Reynolds, 1995], multi-layer perceptrons [Altosaar and Meister, 1995], Radial Basis Functions [Finan et al., 1996] and genetic algorithms [Hannah et al., 1993]

Over the last decade, neural networks have attracted a great deal of attention. They offer an alternative approach to computing and to understanding of the human brain. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. Other advantages include:

Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.

Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage. The objective of this research is to make a comparative study of the existing neural methods for speaker recognition .The speech signal is recorded for number of speakers, and their features are extracted. Feature extraction is done by means of Mel Frequency Cepstrum Coefficients. The neural network is trained by applying these features as input parameters. The features are stored in templates for further comparison. The features for the speaker who has to be identified are extracted and compared with the stored templates using back propagation and radial basis function algorithm. Here, the trained network corresponds to the output; the input is the extracted features of the speaker to be identified. The network does the weight adjustment and the best match is found to identify the speaker. The number of epochs required to get the target decides the network performance.

The paper is organized as follows: in section II, a brief about acoustic processing is presented. In section III, the feature extraction process is revisited. In section IV, the neural approach for speech recognition is reviewed. In section V, a comparative study between the back propagation and radial basis function neural model is presented. In Section VI simulation results are summarized .The conclusion is given in Section VII.

II. ACOUSTIC PROCESSING

Acoustic processing is sequence of processes that receives analog signal from a speaker and convert it into digital signal for digital processing. Human speech frequency usually lies in between 300Hz-8000kHz [2].Therefore 16kHz sampling size can be chosen for recording which is twice the frequency of the original signal and follows the Nyquist rule of sampling [3].The start and end detection of isolated signal is a straight forward process which detect abrupt changes in the signal through a given threshold energy. The result of acoustic processing would be discrete time voice signal which contains meaningful information. The signal is then fed into spectral analyser for feature extraction.

III. FEATURE EXTRACTION

Feature Extraction module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal such that its output eases the work of recognition stage. Main steps involved in feature extraction are explained below:

3.1 Signal Pre Processing

Prior to spectral analysis, there are steps of preprocessing which improve the effectiveness of the feature extraction. Preprocessing may be aimed at noise reduction (one such technique is called spectral subtraction), or at enhancement of formant visibility in the power spectrum. A characteristic of the power spectrum is that higher-frequency formants have lower energy.

Pre-emphasis is a compensatory technique, realized by applying a fixed first order FIR filter with the z-transfer function

$H(z) = 1 - az^{-1}$ ……*(1)*

where ***a*** is the pre-emphasis parameter (usually 0.95).

3.2 Speech Coding

After the captured speech signal is sampled, the utterance is isolated, and the spectrum is flattened, each signal is divided into a sequence of frames, each frame 21ms length and 7ms apart. Then each frame is multiplied by a Hamming window, in order to remove the leakage effects and to smooth the edges[4].

$w(n) = 0.54 - 0.46\cos(2\pi N / n-1)$ ....(2)

where $n \epsilon (0 \text{ to } N-1)$ and $N = 330$.

A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others.
MFCC is perhaps the best known and most popular, and is used in one of the systems implemented in this work.
To simplify the subsequent processing of the signal, useful features must be extracted and the data should be compressed. The power spectrum of the speech signal is the most often used method of encoding. Mel Frequency Cepstral Analysis is used to encode the speech signal. Mel scale frequencies are distributed linearly in the low range but logarithmically in the high range, which corresponds to the physiological characteristics of the human ear. Cepstral analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the speech signal.

3.2.1 Linear Predictive Coding (LPC)

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering and the remaining signal after the subtraction of the filtered modeled signal is called as the residue. The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter (which represents the tube), and run the source through the filter, resulting in speech.

3.2.2 Mel Frequency Cepstrum coefficients (MFCC)

MFCC processor as shown in Fig. 2 analyse speech based on a psychoacoustic modeling which study human auditory perception [5].In MFCC experiment the speech is analyzed with in small fixed size of overlapping windows framed between 20-50 msec .Within this duration the speech signal is considered stationary [6].

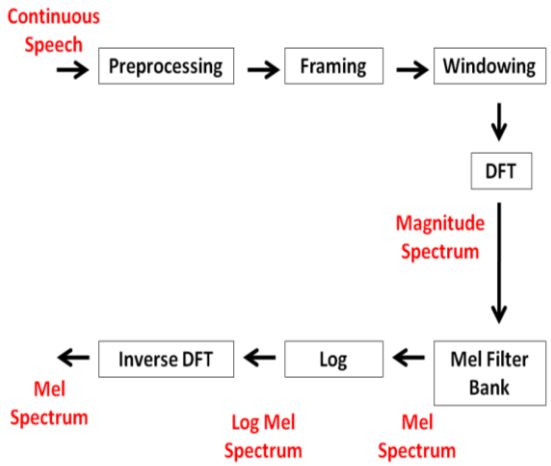

Fig.2. Block diagram of the computation steps of MFCC

The Discrete Fourier Transform of signals is then obtained to get magnitude spectrum of each frame which is calculated as in (3).

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-f\frac{2\pi}{N}kn}$$

..... (3)

where X[k] is the Fourier transform of the signal x[n]. The resulting spectrum is then converted into mel scale using the formula (4):

Mel (f) = 2595*log10 (1+f /700) ..... (4)

where f is the real frequency in hertz.

Take the logs of the powers at each of the mel frequencies. Take the discrete cosine transform which is defined in (5), of the list of mel log powers, as if it were a signal. The MFCCs are the amplitudes of the resulting spectrum.

$$c(n) = \sum_{n=0}^{N-1} \log\left(\left| \sum_{n=0}^{N-1} x[n] e^{-f\frac{2\pi}{N}kn} \right| \right) e^{j\frac{2\pi}{N}b}$$

. ..... (5)

where n = 1, 2,.....∞.

N = number of cepstrum coefficient.

IV. RECOGNIZER

4.1 Neural Approach

The design and function of neural networks simulate functionality of biological brains and neural systems. In the recognition phase, the neural networks are trained to learn the mapping from the features extracted from the pre-separated speech to those extracted from the close-talking microphone speech signal. The outputs of the neural networks are then used to generate acoustic features, which are subsequently used in acoustic model adaptation and system evaluation.

4.1.1 Back Propagation Neural Networks

Back propagation network is one of the most widely used neural networks. It is a multi-layer network which includes at least one hidden layer. First the input is propagated forward through the network to get the response of the output layer. Then, the sensitivities are propagated backward to reduce the error. During this process, weights in all hidden layers are modified. As the propagation continues, the weights are continuously adjusted and the precision of the output is improved.

The partial derivatives of the data function with respect to the parameters of the network are being determined by the back propagation error signals [7].

The equation of error back propagation is

$$\delta j = -\sum_{i \in P_j} \frac{\partial E}{\partial net_i} \frac{\partial net_i}{\partial y_j} \frac{\partial y_j}{\partial net_j}$$

..... (6)

where,

1st factor represents the error of node i;

2nd factor is the weight from unit j to i;

3rd factor is the derivative of node j's activation function.

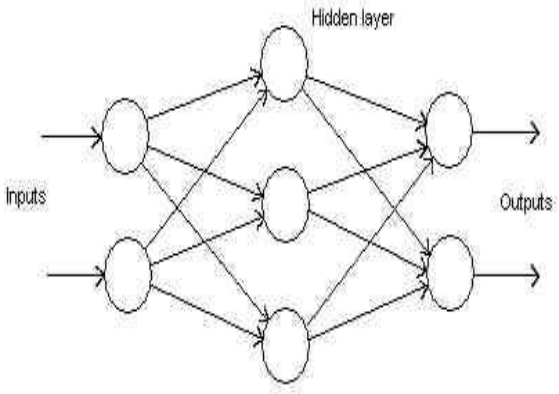

Fig.3. A generalized network

.

Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function which scales the output to a fixed range of values. The output of the limiter is then broadcast to all of the neurons in the next layer. So, to use the network to solve a problem, we apply the input values to the inputs of the first layer, allow the signals to propagate through the network, and read the output values.

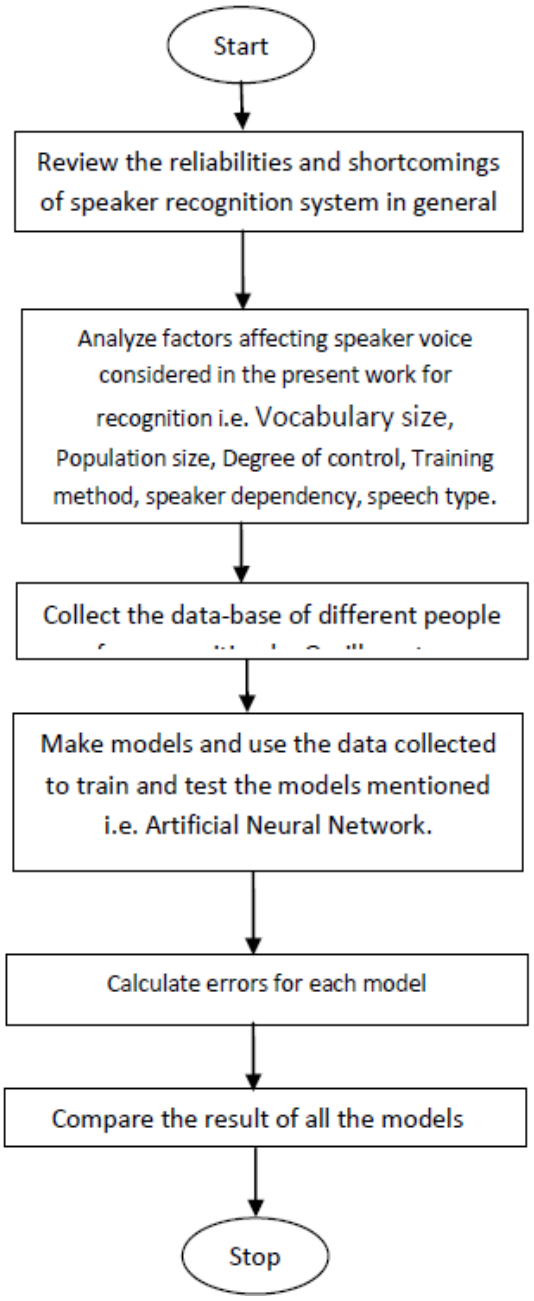

Fig.4. A general flowchart representing the steps involved in simulation

4.1.2 Radial Basis Function (RBF) Neural Networks

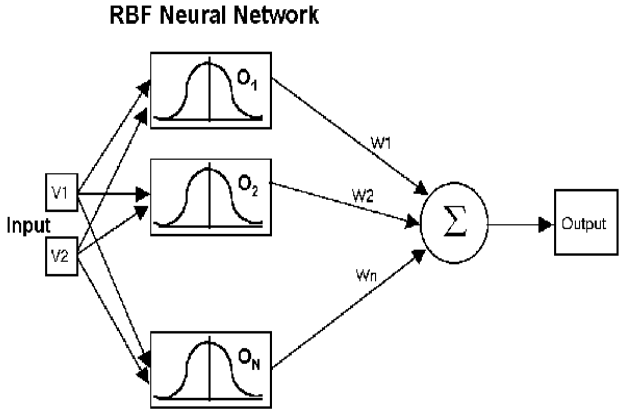

Fig.5. RBF network architecture

RBF networks have three layers:

1.Input layer – There is one neuron in the input layer for each predictor variable. In the case of categorical variables, N-1 neurons are used where N is the number of categories. The input neurons (or processing before the input layer) standardize the range of the values by subtracting the median and dividing by the interquartile range. The input neurons then feed the values to each of the neurons in the hidden layer.

2.Hidden layer – This layer has a variable number of neurons (the optimal number is determined by the training process). Each neuron consists of a radial basis function centered on a point with as many dimensions as there are predictor variables. The spread (radius) of the RBF function may be different for each dimension. The centers and spreads are determined by the training process. When presented with the x vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function to this distance using the spread values. The resulting value is passed to the the summation layer.

3. Summation layer – The value coming out of a neuron in the hidden layer is multiplied by a weight associated with the neuron (W1, W2, ...,Wn in this figure) and passed to the summation which adds up the weighted values and presents this sum as the output of the network. Not shown in this figure is a bias value of 1.0 that is multiplied by a weight W0 and fed into the summation layer. For classification problems, there is one output (and a separate set of weights and summation unit) for each target category. The value output for a category is the probability that the case being evaluated has that category.

Training RBF Networks

The RBF network is a three-layer feed-forward neural network, between the input and the output layers there is a "hidden layer". When training, vectors are input to the first layer and fanned out to the hidden layer. In the latter, a cluster of radial basis functions turn the input to output, adjusting the weight of the input to the hidden layer. Then, under the target vector's supervising, the weight of the output vector of the hidden layer is adjusted. When clustering texts, the Euclidean Distance between the input vectors and the weight vectors, which have been adjusted by training process, is calculated. Each input sample is sorted to a class. Then the output layer collects samples belonging to same classes and organizes an output vector, the final clustering. Radial basis networks can require more neurons than standard feed forward back propagation networks, but often they can be designed in a fraction of the time it takes to train standard feed forward networks. In the network, n–dimensional input feature vector are accepted to the input layer which is a set of n units consisting of input vectors x1, x2 ... xn and output of input layer is input to the hidden layer. The output of the hidden layer is then multiplied by the weighting factor w (i, j) which is the input to the output layer of the network y (x).

$$y(x) = \sum_{i=1}^{N} w_i \Phi(\| x - c_i \|)$$

..... (7)

where the approximating function y(x) is represented as a sum of N radial basis functions, each associated with a different centre c_i, and weighted by an appropriate coefficient w_i, and || || indicates the Euclidean norm on the input space [8].

V. SIMULATION

In this paper, we study and explore the use of neural networks in the authentication of speaker, using the features extracted from his voice and then comparing them with neural models and then calculating the results.

In the neural models made in the paper four input factors have been taken. These factors are peak frequency, peak amplitude, total power and signal to noise ratio (SNR). These factors have been extracted with the help of the software 'Oscillometer'.

5.1 Back Propagation Neural Network Model

Input data has been taken with the help of software by taking the different person's speech. A backpropagation network is made on MATLAB software with the statistical details given in Table1 .

Table 1: Statistical Data for BP neural model

Neurons in input layer	4
Number of hidden layer and number of neurons in hidden layer	1,4
Neurons in output layer	1
Transfer function (input, hidden and output)	Tansigmoid, tansigmoid, linear
Epochs	1400

Expaph of mean square error (MSE) with epochs has been shown. We can conclude from the graph that error reduces with epochs.

5.2 Radial Basis Function Neural Network Model

Input data taken in the earlier BP model is taken here also.A Radial Basis Function network is made again with MATLAB software with the statistical details given in Table

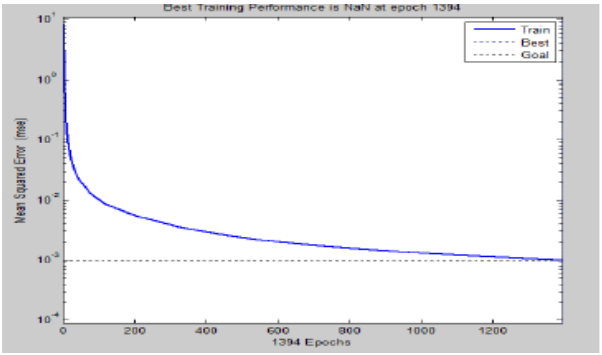

Fig. 6. Learning with BP neural network model

Table 2: Statistical Data for RBF neural model

Neurons in input layer	4
Number of radial basis layer and number of neurons in hidden layer	1,7
Neurons in output layer	1
Transfer function (input, hidden and output)	Tansigmoid, tansigmoid, linear
Epochs	50

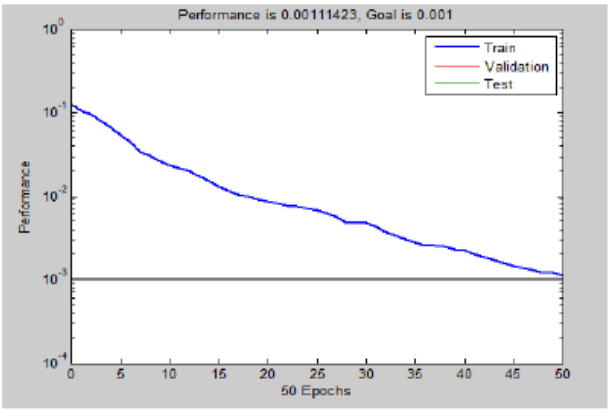

Fig. 7. Learning with RBF neural network model.

In this figure a graph of MSE with epochs has been shown and it can be seen that error reduces with epochs.

VI. RESULTS

Experimental output results for both the models i.e. RBF and BP neural network models have been calculated. Results show that the recognition performance for a particular speaker's speech is better recognised with RBF model than the BP model.

The results for the recognition test with BP and RBF models are shown in Table3.

Table 3: Results

Models	BP	RBF
Recognition rate	78.1%	81.1%

It can be seen from the above table that RBF model has more recognition rate than BP model. Hence, we can say that RBF neural model is better than BP neural model.

VII. CONCLUSIONS

It can be seen from the above experimental results that among the two classifiers mentioned above i.e. RBF and BP neural models, RBF overrules BP model. RBF model is able to recognize more number of speakers than recognised by BP model. Hence, RBF model is better than BP model.

Secondly, it can be seen that computational time taken by RBF is much lesser than BP. The RBF network based models are linear in the parameters and therefore guarantee convergence to their optimum values for particular network architecture. Development of the RBF network model therefore requires less trial and error and thus, less time and effort, than that needed by the MLP with BP approach.

REFERENCES

1. Md Sah Bin Hj Salam, Dzulkifli Mohamad Sheikh Hussain Shaikh Salleh," Temporal Speech Normalization Methods Comparison in Speech Recognition Using
   Neural Network", International Conference of Soft Computing and Pattern Recognition, 2009
2. Khalifa, O.O, et. al, "Speech coding for Bluetooth with CVSD algorithm", Proc. RF and Microwave Conference. Selangor, Malaysia, Page(s):227 – 229, 5-6 Oct. 2004
3. Young, S., "A review of large vocabulary continuous speech", IEEE Signal Processing Magazine, v. 13, n 5, pp 45-57, 1996
4. Anup Kumar Paul1, Dipankar Das2, Md. Mustafa Kamal3," Bangla Speech Recognition System using LPC and ANN",Seventh International Conference on Advances in Pattern Recognition,2009
5. Premakanthan, P.; Mikhael, W.B., "Speaker verification/recognition and the importance of selective feature extraction: review", Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems, 2001. MWSCAS 2001. Volume 1, 14-17 Page(s):57 –61. Aug. 2001
6. Parson, T.W, "Voice and Speech Processing", New York, United States of America: McGraw-Hill. 294, 1987.
7. Hui Kong, Xuchun Li, Lei Wang, Earn Khwang Teoh, Jian-Gang Wang, Venkateswarlu, R "Generalized 2D principal component analysis", Proc. 2005 IEEE International Joint Conference on Volume 1, Aug. 2005.
   Harry Wechsler, Vishal Kakkad, Jeffrey Huang, Srinivas
   Gutta, V. Chen, "Automatic Video-based Person Authentication Using the RBF Network" First International Conference on Audio- and Video-Based Biometric Person Authentication, 1997 pages 85-92.
8. Gabriel Zigelboim and Dr Ilan D. Shallom," A comparison Study of Cepstral Analysis with Applications to Speech Recognition", International Conference on Information Technology: Research and Education,2006
9. Yongjin Wang and Ling Guan, "an investigation of speech-based human emotion recognition", IEEE 6th Workshop on Multimedia Signal Processing, 2004