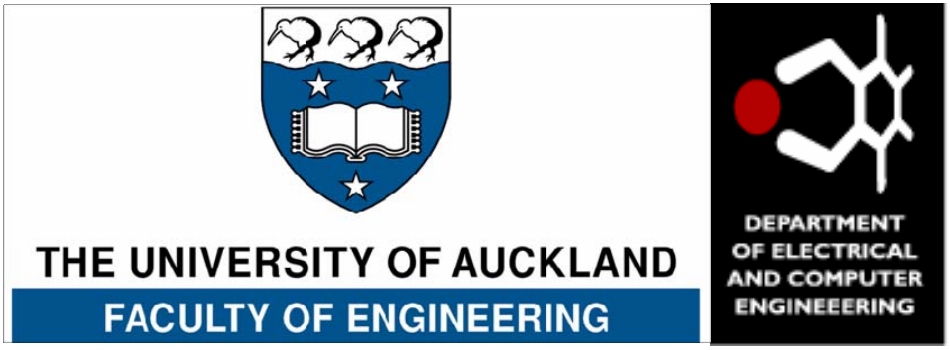

Gammatone Auditory Filterbank and Independent Component Analysis for Speaker Identification systems

School of Engineering Report No. 622

Yushi[1] Zhang and Waleed[2] H. Abdulla
Department of Electrical and Computer Engineering
The University of Auckland, Private Bag 92019,
Auckland, New Zealand
[1]yzha104@ec.auckland.ac.nz [2]w.abdulla@auckalnd.ac.nz

2005

ABSTACT

Speaker identification is the process of recognizing who is speaking on the basis of information extracted from the speech signal. It has a number of applications in security and voice controlled service area. However, the most commonly used speaker recognition techniques work successfully only in clean or matched environment. Accurate speaker identification is made difficult due to a number of factors, with handset/channel mismatch and environmental noise being two of the most prominent. This paper presents a new novel technique which is based on the Gammatone filterbank (GTF) and independent component analysis (ICA). Compared with some other standard techniques, the new technique holds the best performance for a speaker text-independent identification system with mismatched environment and low environmental SNR level (less than 20dB).

Key-words: Gammatone filterbank, independent component analysis, speaker identification

1. INTRODUCTION

Speaker identification is a task of determining the best-matching speaker for an unknown speaker from a database of known speakers. The ability to identify people uniquely by their voice alone leads to several applications; such as in law enforcement, speaker identification system can be used to help identify suspects with their recorded voice. Security application area; access to cars, buildings, bank accounts and other services may be voice controlled. And in meetings, conference, or conversations, the identification system makes machine identification of participants possible. If used in conjunction with continuous speech recognition systems, transcription could be produced containing a record of who said what [1, 2].

There are a number of techniques commonly used in speaker identification system; linear predictive cepstral coefficients (LPCC) derived from LP model divided the log magnitude spectrum of a speech signal into two components: the excitation source component is a fast changing component and the vocal tract component is a slow varying component [8]. By taking inverse Fourier transform, the vocal tract component is resided in the lower order cepstral coefficients. Mel-frequency cepstral coefficients (MFCC) uses filter to approximate the human auditory system. The centre frequencies of filterbank follow the mel-frequency scale which is linear up to 1000 Hz and logarithmic above 1000Hz [7]. Perceptual Linear Predictive (PLP), which first used the concepts from the psychophysics of the hearing to derive the estimate of the auditory spectrum [3]. In PLP technique, several well known properties of hearing are simulated by practical engineering approximations: such as critical-band analysis, equal-loudness pre-emphasis, and intensity-loudness conversion. And the resulting auditory-like spectrum of speech is approximated by the autoregressive all-pole model. Those techniques have been proved that they worked well in a clean or match environment. However, the mismatch in acoustic characteristics between speech signals produced by the training speakers and those by the testing speakers has been causing serious performance degradation for speaker identification system. Meanwhile, modern speech processing applications require operation on signal of

interest that is contaminated by high level of noise. These situations call for a greater robustness in estimation of the speech parameters for mismatch environment and low environmental SNR level.

In this paper, we propose an algorithm, which efficiently remove the effect of mismatch environment and environmental noise. The proposed method is based on the Gammatone filterbank (GTF) and independent component analysis (ICA). Gammatone filterbank modeling is a physiologically based strategy followed in mimicking the structure of the peripheral auditory processing stage [4]. It models the cochlea by a bank of overlapping bandpass filters [4]. And ICA has been shown a highly effective in extracting the features from the given set of observed speech signals [6], by reflecting the statistical structure of the observed signal. Thus the features extracted by ICA emphasis the difference of the statistical structures among the speakers, which can model the distribution of the individual speaker [5]. The proposed method was compared with three commonly used techniques; LPCC, MFCC and PLP in a text-independent speaker identification system on 100 speakers form the TIMIT database. The simulation results prove that the proposed features are more robust to mismatch environment with low environmental SNR level.

The paper is organized as following: section 2 introduces the Gammatone auditory filterbank. Then ICA is depicted in section 3. After that, the new method based on GTF and ICA is discussed in section 4. Section 5 demonstrates simulations tested with new method. Finally, conclusions are given in section 6.

2. GAMMATONE AUDITORY FILTERBANK

The digital filter bank is one of the most fundamental concepts in speech processing. In auditory modelling, filterbank resembles the characteristics of the basilar membrane (BM). In the inner ear's cochlea, the input speech signals induce mechanical vibration on the basilar membrane. And each position of basilar membrane responds to some localized frequency information of the speech signals. Then in auditory modelling, each bandpass filter is modelled by these frequency characteristics of basilar membrane [6]. In this section, Gammatone auditory filterbank (GTF) is discussed briefly.

2.1 Gammatone Filter

Gammatone auditory filter banks are non-uniform overlap bandpass filters, designed to imitate the frequency resolution of human hearing [4]. The impulse response of each filter follows the Gammatone function shape. This function has the following classical form [4]:

$$h(t) = \gamma(n,b)t^{n-1}e^{-bt}\cos(\omega t + \phi)u(t) \qquad (2.1)$$

where, $\gamma(n,b)$ is a normalization constant depending on the order, *n*, and the bandwidth related factor, *b*. *w* is the radian centre frequency. ϕ is the phase shift and *u(t)* is a unit step function. The Gammatone function corresponding to a cochlea filter (order is 4) centered at 1000Hz and with bandwidth of 125Hz is shown in Fig.1.

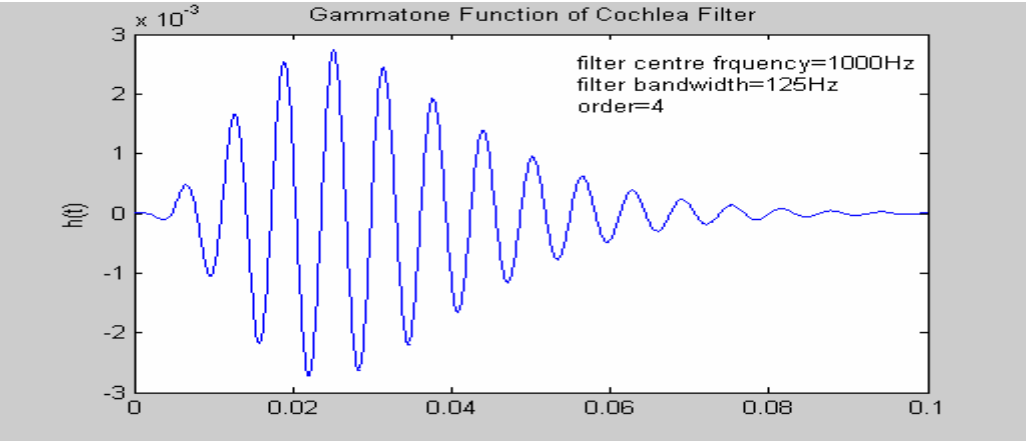

Figure 1. The Gammatone Function

2.2 Bandwidth and Centre Frequency of GTF

The bandwidth of each filer in the Gammatone filterbank is determined according to the auditory critical band (CB) corresponding to its centre frequency. The CB is the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea path [4]. The bandwidth can be approximated as following formula at centre frequency f_c:

$$b = 24.7(1 + 4.37 f_c) \qquad (2.2)$$

Thus to determine the bandwidth of each filter, the centre frequency of each filter has to be ready beforehand. In the human auditory system, there are around 3000 inner hair cells along the 35mm spiral path cochlea. Each hair cell could resonate to a certain frequency within a suitable critical bandwidth. This means that there are approximately 3000 bandpass filters in the human auditory system. This resolution of filters can not be implemented practically using computational modeling techniques. However we can approximate his high resolution into some possibly implemented one. This can be achieved by specifying certain overlapping between the contiguous filters. The percentage-overlapping factor, *v*, will specify the number of channels, filters, required to cover the useful frequency band [4]. If we suppose that the information carrying band is bounded by f_H Hz and f_L Hz with *v* overlapping spacing the number of filters will be:

$$N = \frac{9.26}{v} \ln \frac{f_H + 228.7}{f_L + 228.7} \qquad (2.3)$$

Then the centre frequency can be calculated by

$$f_c = -228.7 + (f_H + 228.7) e^{-\frac{vn}{9.26}} \qquad (2.4)$$

where $1 \le n \le N$

Having decided the location of the centre frequency of each filter, the bandwidth can be calculated from (2.2) and we can now proceed to the implementation stage.

2.3 Gammatone Filter Implementation

The previous sections described a physiologically motivated way for deciding the bandwidth and the center frequency of each filter in the Gammatone filterbank. The implantation of a bandpass filter from its time domain function is a straightforward procedure in signal processing [4]. It is simply started by finding the Laplace transform of the Gammatone function, then mapping it into the digital form using bilinear transform or impulse invariant transform. The frequency impulse response of a 16-channel filterbank, covering 100-8000Hz band, is shown in Fig 2.

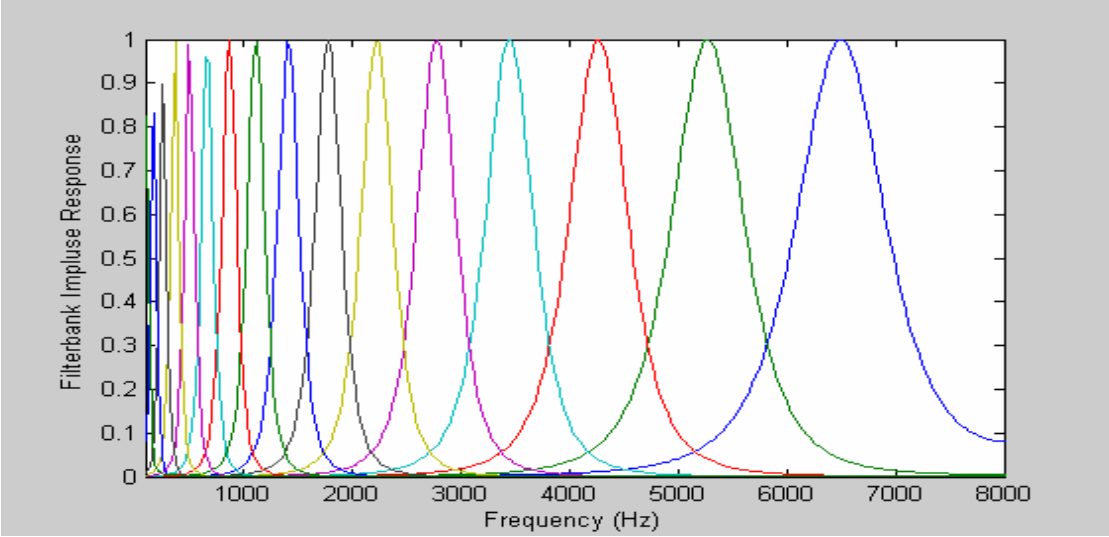

Figure 2. Frequency Response of a Gammatone Filterbank

The bandwidth of the channel is logarithmically proportional with the centre frequency. Thus, GTF can very well model the non-linear frequency characteristics of the cochlea even it is belonging to the linear system family [4].

3. INDPEDENT COMPOONENT ANALYSIS

Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate (multidimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both statistically independent, and nongaussian [9]. Here we briefly introduce the basis concepts, and estimation principles of ICA.

3.1 Basic Definitions

Consider a situation where there are a number of signals emitted by some physical objects or sources. Assume that there are several sensors or receivers. These sensors are in different positions, so that each records a mixture of the original source signals with slightly different weights. For the sake of simplicity of exposition, let us say there are three underlying source signals, and also three observed signals. Denote by $x_1(t)$, $x_2(t)$ and $x_3(t)$ the observed signals, and by $s_1(t)$, $s_2(t)$ and $s_3(t)$ the original source signals. The $x_i(t)$ are then weighted sums of the $s_i(t)$, where the coefficients depend on the distances between the sources and the sensors:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) &= a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{aligned} \qquad (3.1)$$

The a_{ij} are constant coefficients that give the mixing weights. They are assumed unknown, since we cannot know the values of a_{ij} without knowing all the properties of the physical mixing system. The source signals s_i are unknown as well, since the very problem is that we cannot record them directly.

What we would like to do is to find the original signals from the mixtures $x_1(t)$, $x_2(t)$ and $x_3(t)$. We can safely assume that the mixing coefficients a_{ij} are different enough to make the matrix that they form invertible. Thus there exists a matrix **W** with coefficients w_{ij}, such that we can separate the s_i as:

$$
\begin{aligned}
s_1(t) &= w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\
s_2(t) &= w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\
s_3(t) &= w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t)
\end{aligned} \qquad (3.2)
$$

Such a matrix **W** could be found as the inverse of the matrix that consists of the mixing coefficients a_{ij} in Eq.3.1, if we knew those coefficients a_{ij}.

The question now is: how can we estimate the coefficients w_{ij} in Eq.3.2? All we observe is the signals x_1, x_2 and x_3, and we want to find a matrix W so that the representation is given by the original source signals s_1 s_2 and s_3.

A surprisingly simple solution to the problem can be found by considering just the statistical independence of the signals. In fact, if the signals s_1 s_2 and s_3 are independent, then they can be recovered (They could be multiplied by some scalar constants, though, but this has little significance) [9]. Thus the problem boils down to finding a linear representation in which the components are statistically independent. In practical situations, we cannot in general find a representation where the components are really independent, but we can at least find components that are as independent as possible [9].

This leads us to the following definition of ICA, given a set of observations of random variables $(x_1(t), x_2(t), ..., x_n(t))$, where *t* is the time or sample index, assume that they are generated as a linear mixture of independent components:

$$
\begin{pmatrix} x_1(t) \\ x_2(t) \\ \cdot \\ \cdot \\ \cdot \\ x_n(t) \end{pmatrix} = A \begin{pmatrix} s_1(t) \\ s_2(t) \\ \cdot \\ \cdot \\ \cdot \\ s_n(t) \end{pmatrix} \qquad (3.3)
$$

where **A** is some unknown matrix. ICA now consists of estimating both the matrix **A** and the $s_i(t)$, when we only observe the $x_i(t)$.

Alternatively, we could define ICA as follows: find a linear transformation given by a matrix **W,** so that the random variables $s_i(t)$, i=1,…,n are as independent as possible.

3.2 Methods for ICA

In this section, we introduce some principles for estimating the model of independent component analysis (ICA).

3.2.1 ICA by maximization of kurtosis

This method is based on the central theorem given following [10]:

If a set of signals $S = (s_1, s_2, ..., s_{M)}$ are independent with means ($\mu_1, \mu_2, ..., \mu_M$) and variances ($\sigma_1^2, \sigma_2^2, ..., \sigma_M^2$) then, for a large number *M* of signals *S*, the signal

$$x = \sum_{j=1}^{M} s_j$$

has a pdf which is approximately Gaussian with mean $\sum_j \mu_j$ and variance $\sum_j \sigma_j^2$.

Now the ICA problem becomes the problem: how to find a matrix W in Eq 3.2 to make the separated signals s_i have maximum nongaussianity? Thus we must have a quantitative measure of nongaussianity of a random variable, say *y*. Kurtosis has been widely used as a measure of nongaussianity in ICA and related field. The kurtosis of *y*, denoted by *kurt(y)*, is defined by

$$kurt(y) = E\left[y^4\right] - 3(E\left[y^2\right])^2 \qquad (3.4)$$

where $E[.]$ denotes the expected values. Typically nongaussianity is measured by the absolute value of kurtosis. The measure is zero for a Gaussian variable, and greater than zero for most nongaussian random variables.

In practice, to maximize the absolute value of kurtosis, we would start from some vector **W**, compute the direction in which the absolute value of the kurtosis of $y = w^T z$ is growing most strongly, where *T* denotes the transpose (we take the transpose because all vectors in this paper are column vectors), base on the available sample $z(1),...z(T)$ of mixture vector **z** (**z** is new vector transferred from observed data vector **x** by whitening), and then move the vector **W** in that direction. This idea is implemented in gradient methods and their extensions. It introduces a fast fixed-point

algorithm using kurtosis discussed in table 3.1. The great detail about derivation can be seen in [9].

1. center the data to make its mean zero.
2. whiten the data to give **z**.
3. choose an initial (e.g. random) vector w of unit norm.
4. $w \leftarrow E\left[z(w^T z)^3\right] - 3w$.
5. let $w \leftarrow w/\|w\|$.
6. if not converged, go back to step 4.

Table 3.1 the fastICA algorithm for finding maximizing kurtosis

3.2.2 ICA by maximization of negentropy

Negentropy is the second important measure of nongaussianity. Its properties are in many ways opposite to those of kurtosis: it is robust. Negentropy is based on the information-theoretic quantity of differential entropy. The entropy of a random variable is related to the information that the observation of the variable gives. The more “random”, i.e., unpredictable and unstructured the variable is, the larger its entropy [9]. The entropy ***H*** of a random vector *y* with density $p_y(\eta)$ is defined as

$$H(y) = -\int p_y(\eta) \log p_y(\eta) d\eta \qquad (3.5)$$

A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance. To obtain a measure of nongaussianity that is zero for a Gaussian variable and always nonnegative, one often uses a normalized version of differential entropy, called negentropy. Negentropy ***J*** is defined as follows

$$J(y) = H(y_{gauss}) - H(y) \qquad (3.6)$$

where y_{gauss} is a Gaussian random variable of the same correlation (and covariance) matrix as *y*.

In practice, negentropy is approximated by

$$J(y) \propto \{E[G(y)] - E[G(v)]\}^2 \qquad (3.7)$$

where G can be any nonpolynomial functions. The following choices of G have proved very useful [9]:

$$G_1(y) = \frac{1}{a_1}\log\cosh a_1 y \qquad 1 \le a_1 \le 2 \qquad (3.8)$$

$$G_2(y) = -\exp(-y^2/2) \qquad (3.9)$$

$$G_3(y) = y^4 \qquad (3.10)$$

Their derivatives respecting to y are denoted as following

$$g_1(y) = \tanh(a_1 y) \qquad 1 \le a_1 \le 2 \qquad (3.11)$$

$$g_2(y) = y\exp(-y^2/2) \qquad (3.12)$$

$$g_3(y) = y^3 \qquad (3.13)$$

As with kurtosis, we can derive a simple gradient algorithm for maximizing negentropy. Taking the gradient of the approximation of negentropy in Eq.3.7 with respect to **w**, and taking the normalization $E[(w^T z)^2] = \|w\|^2 = 1$ into account, one obtains the following fast fixed-point algorithm. The great detail about derivation can be seen in [9].

---

1. center the data to make its mean zero.
2. whiten the data to give **z**.
3. choose an initial (e.g. random) vector w of unit norm.
4. let $w \leftarrow E[zg(w^T z] - E[g'(w^T z)]w$, where g is defined as in Eq. 3.11-Eq.3.13
5. let $w \leftarrow w/\|w\|$.
6. if not converged, go back to step 4.

Table 3.2 the fastICA algorithm for finding maximizing negentropy [9]

3.2.3 ICA by maximum likelihood estimation

A very popular approach for estimating the ICA model is maximum likelihood estimation. The density p_x of the mixture vector

$$X = AS \tag{3.14}$$

can be formulated as

$$p_x(X) = |\det W| p_s(S) = |\det W| \prod_i p_i(s_i) \tag{3.15}$$

where $W = A^{-1}$, and the p_i denote the densities of the independent components. This can be expressed as a function of $W = (w_1, ..., w_n)^T$ and **x**, giving

$$p_x(X) = |\det W| \prod_i p_i(w_i^T x) \tag{3.16}$$

Assume that we have T observations of **x**, denoted by $x(1), x(2), ..., x(T)$. Then the likelihood can be obtained as the product of this density evaluated at the T points. This denoted by L and considered as a function of W:

$$L(W) = \prod_{t=1}^{T} \prod_{i=1}^{n} p_i(w_i^T x(t)) |\det W| \tag{3.17}$$

Very often it is more practical to use the logarithm of the likelihood, since it is algebraically simpler. And after a simplification, Eq.3.17 can be converted to

$$\frac{1}{T} \log L(W) = E\left[\sum_{i=1}^{n} \log p_i(w_i^T x) \right] + \log |\det W| \tag{3.18}$$

To perform maximum likelihood estimation in practice, a natural gradient algorithm is used and introduces a fast fixed-point algorithm seen in table 3.3. The great detail about derivation can be seen in [9].

---

1. center the data to make its mean zero.
2. whiten the data to give **z**.
3. compute correlation matrix $C = E[xx^T]$.
4. choose an initial (e.g., random) separating matrix W.
5. compute

$$y = Wz$$

$$\beta_i = -E[y_i g(y_i)], \quad \text{for } i=1,\ldots,n$$

$$\alpha_i = -1/(\beta_i + E[g'(y_i)]), \quad \text{for } i=1,\ldots,n \text{ and g is the tanh function}$$

6. update the separating matrix by

$$W \leftarrow W + diag(\alpha_i)\{diag(\beta_i) + E[g(y)y^T]\}W$$

7. decorrelate and normalize by

$$W \leftarrow (WCW^T)^{-1/2}W$$

8. if not converged, go back to step 3.

Table 3.3 the fastICA algorithm for maximum likelihood estimation

3.2.4 other methods for ICA

Except the methods mentioned above, there are a number of other methods available:

An estimation principle for ICA that is very closely related to maximum likelihood is the infomax principle. This is based on maximizing the output entropy, or information flow, of a neural network with nonlinear outputs. Hence the name infomax [9].

One approach for ICA estimation, inspired by information theory, is minimization of mutual information, since mutual information is a natural measure of the dependence between random variables [9].

Another important approach for ICA consists of using higher-order cumulant tensor, especially, for the fourth-order cumulant tensor [9].

3.3 Extensions of ICA

In this section, we will introduce some extensions of ICA.

Noisy ICA: In real life, there is always some kind of noise present in the observations. Noise can correspond to actual physical noise in the measuring devices, or to inaccuracies of the model used. Therefore, it has been proposed that the ICA model should include a noise term as well.

ICA with Overcomplete Bases: a difficult problem in ICA is encountered if the

number of mixtures x_i is smaller than the number of independent components s_i. This means that the mixing system is not invertible. Therefore, even if we knew the mixing matrix exactly, we could not recover the exact values of the independent components.

Nonlinear ICA: in many situations, the basic linear ICA is too simple for describing the observed data **x** adequately. Hence, it is natural to consider extension of the linear model to nonlinear mixing models.

Since these problems are out of our research scope, they will not be discussed in this paper. The great detail about these extensions and related methods can be seen in [9].

4. FEATURE MATRIX EXTRACTED USING GTF AND ICA

In this section a new feature extraction method for speaker identification is discussed. This method is based on Gammatone filterbank and Independent Component Analysis. We begin with introducing the method. After that, we discuss how to combine the new method with speaker identification system. Finally, we introduce an optimized method to improve the performance for speaker identification.

4.1 GTF-ICA Feature Matrix

A new feature extraction method based on GTF and ICA for speaker identification is developed. Compared with the commonly used techniques, the feature extracted using proposed method is a matrix instead of the vector. The flow chart of the GTF-ICA feature matrix calculation algorithm is shown in Fig 3.

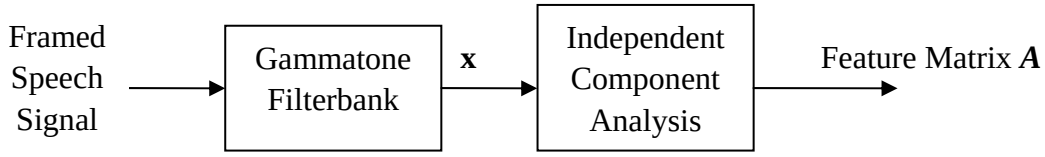

Figure 3 The algorithm of GTF-ICA feature matrix extraction

From the above figure, the GTF-ICA feature matrix is extracted as following:

Firstly, the framed speech signal passes to a Gammatone filterbank, assume the number of filters is N. Thus the output of this filterbank is a matrix **x**, which has N rows, and each row x_i, i=1,…,N, represents the output of each bandpass filter of Gammatone filterbank in time domain.

After that, independent component analysis is done to the matrix **x**. Recall the ICA formula in section 3.1:

$$\begin{aligned} x_1 &= a_{11}s_1 + a_{12}s_2 + \ldots + a_{1N}s_N \\ x_2 &= a_{21}s_1 + a_{22}s_2 + \ldots + a_{2N}s_N \\ &\vdots \\ x_N &= a_{N1}s_1 + a_{N2}s_2 + \ldots + a_{NN}s_N \end{aligned} \qquad (4.1)$$

Here, we assume that the framed speech signal is linearly combined with N basis signals $s_{i,}$, i=1,…,N. This assumption is valid for all kind of signals. Thus, we believe that coefficients $a_{i1}, a_{i2}, ..., a_{ii}$, i=1,…,N, reflect the statistical structure of the framed signal in a particular frequency band. And this particular frequency band corresponds the ith filter of Gammatone filterbank. Therefore, the whole matrix $\boldsymbol{A}$ is a $N \times N$ matrix with coefficients a_{ij}, and reflects the statistical structure of the framed speech signal in different frequency bands which are designed to imitate the frequency resolution of human hearing. And since ICA leads a highly efficient representation of the observed data, we believe that the feature matrix $\boldsymbol{A}$ extracted from the speech signal may focus on the difference of the statistical structures among the speakers. Hence it can be used to model the distribution of the individual speaker.

4.2 GTF-ICA Feature Matrix and Speaker Identification System

In this section, GTF-ICA feature matrix is used for a speaker text-independent identification system. Fig 4 shows a speaker identification system using GTF-ICA feature matrix. The main modules of the system are discussed following:

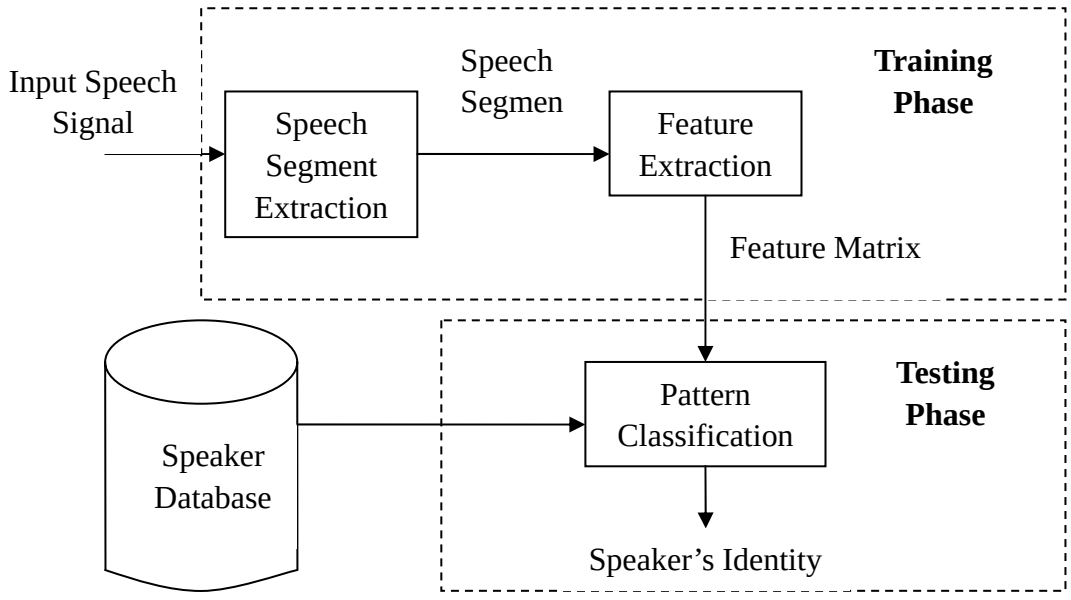

Figure 4 The block diagram of a speaker identification system using GTF-ICA feature matrix

4.2.1 speech segment extraction

For using ICA, normally, speech signal do not need to be segmented, since ICA technique can deal with any kinds of signal, no matter the signal is stationary or not. However, in our experiments, we prefer to segment the signal, because more segments we take, more feature matrix we generate, and more accuracy distribution of speaker we can obtain. However, for a certain long signal, more segments means shorter segment length. For a short segment, the feature matrix extracted using ICA may emphasis the difference of the linguistic information instead of the difference of the statistical structure among speakers. We determined the segment length experimentally as 3 seconds with 96.67% overlapping.

4.2.2 feature extraction

In this module GTF-ICA feature matrix of each segment is extracted using the method discussed in section 4.1, the mean of these feature matrixes is taken, and stored as a distribution of a speaker. In this paper we used maximizing negentropy method mentioned in section 3.2.2 to estimate the feature matrix during ICA, and we choose the nonpolynomial function as $g(y) = \tanh(a_1 y)$, $1 \le a_1 \le 2$.

There is a parameter in this module: the number of filters. In this paper, we determined this value experimentally, and found 30 is the best choice.

4.2.3 pattern classification

It decides the speaker's identity based on the feature matrix extracted from the target speaker's speech signal and the speaker distribution stored in the database. In this module, to reduce the computation load, we used whole testing sentence to generate a feature matrix instead of segmenting the signal and taking the mean of feature matrixes extracted from segments.

Thus, for speaker identification system, each speaker is presented as a training feature matrix generated using her/his training sentences. Then the testing feature matrix extracted from the target speaker is compared with these training feature matrixes to

find the most similar training feature matrix. Then, it is considered that the speaker whom is represented by this training feature matrix has the same identity as the target speaker has.

Normally, there are two techniques for measuring the similarity of two matrixes: matrix distance measure and correlation measure. Huge simulations have been done that prove the correlation measure method is an efficient and reliable method for our system. The correlation of two matrixes can simply be calculated using Matlab function “corrcoef”. The score lies between 0 and 1. The bigger score means the bigger similarity between two matrixes.

4.3 Optimized GTF-ICA Feature Matrix

Since the GTF-ICA feature matrix reflects the statistical structure of the speech signal in different frequency bands and denotes the distribution of the individual speaker, we believe that the dynamic coefficients of matrix may also be useful for speaker identification. An optimized GTF-ICA method for speaker identification is developed as following:

For a GTF-ICA feature matrix $\boldsymbol{A}$, with coefficients a_{ij}, $1 \le i, j \le N$, its two delta matrixes $\boldsymbol{B}$ and $\boldsymbol{C}$ are generated from the feature matrix.

For row delta matrix $\boldsymbol{B}$ with size $(N-1)\times N$, its coefficient b_{ij} is calculated according Eq.4.1:

$$b_{ij} = a_{ij} - a_{i-1,j}, \qquad 1 \le i \le N-1,\ 1 \le j \le N \tag{4.1}$$

For column delta matrix $\boldsymbol{C}$ with size $N \times (N-1)$, its coefficient c_{ij} is calculated according Eq.4.2:

$$c_{ij} = a_{ij} - a_{i,j-1}, \qquad 1 \le i \le N, \quad 1 \le j \le N-1 \tag{4.2}$$

Therefore, for each matrix, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, we put them into pattern classification module and obtain three scores, $score_A$, $score_B$, and $score_C$. Then a total score is obtained by adding up these scores with multiplying some weights. The formula can be seen in Eq.4.3:

$$totalscore = p_A \times score_A + p_B \times score_B + p_C \times score_C \qquad (4.3)$$

where p_A, p_B and p_C are some weights can be determined experimentally.Then the higher score holds the higher similarity of two matrixes. A number of simulations were carried on to determine values of weights, $p_A = 0.58$, $p_B = 0.33$ and $p_C = 0.09$.

5. EXPERIMETAL EVALUATION

In this section, we try to simulate a real situation for speaker identification. Normally, in training phase, the speech signal for training could be record in a clean environment or made to be clean by using some speech enhancement techniques. However, in test phase, it is impossible to make sure that the speech signal from the target speaker is clean. There are usually some background noises when the target speaker talks to the identification machine, and no time to reduce noises since we need an immediate respond from the machine.

This leads to a serious problem in modern speaker identification system: mismatch environment with low environmental SNR level. The commonly used techniques fail in this situation. In this section, we will investigate the effect of additive noise for our new technique: GTF-ICA.

5.1 Database Description

The experiments were primarily conducted using the TIMIT and NOISEX-92 speech databases.

TIMIT is a noise free speech database recorded using a high quality microphone sampled at 16Hz. In this paper, 100 speakers from 8 dialects in the testing and training folder of TIMIT were used in this experiment. In TIMIT, each speaker produces 10 different sentences. The average duration of each sentence is about 3 seconds. In my experiment the 1^{st}, 3^{rd}, 5^{th}, 7^{th}, 9^{th} and 10^{th} sentences were used for training and the remaining sentences were used for testing. In our experiments, the total length of training sentence is about 18 seconds and is about 5 seconds for testing sentence.

NOISEX-92 is standard noise database. It provides various military noises recorded in real environments with sampling rate: 19.98 KHz.

In our experiments, the clean TIMIT training sentence from each speaker was used to generate the GTF-ICA feature matrix. And these matrixes were stored in the database and denoted each speaker. After that, their testing sentences added with noises taken from NOISEX-92 (noise was firstly down-sampled to 16Hz) were used to generate

feature matrixes. And we compared them with feature matrixes stored in the database to identify whom the test sentence belonged to. The identification rate was defined as following equation:

$$identification\ rate = \frac{number\ of\ the\ correctly\ identified\ speakers}{the\ total\ number\ of\ speakers\ used\ to\ be\ identified} x100\% \quad (5.1)$$

5.2 Experiments Description

The GTF-ICA feature matrix and its optimized method were tested in a speaker identification system with mismatch environment and low environmental SNR level. Simulation results were compared with some commonly used feature extraction techniques' performances. Here we used LPCC, MFCC and PLP with 24 orders and 32 components Gaussian Mixture Models (GMM) [11] as baseline techniques.

5.2.1 for white Gaussian noise

The white Gaussian noise was added on the testing sentences to simulate the situation where the target speaker talked to the identification machine in a Gaussian noise background environment. Fig.5 shows the simulation results:

5.2.2 for actual noise

In this section, some real noises were used instead of Gaussian white noise in the last section. The noises were taken from the NOISEX-92 noise database. There are some brief descriptions about the noises we used.

'detoryerops' is a destroyer operations room background noise.

*'**buccaneer**'* is the noise recorded at a buccaneer jet, which was moving at a speed of 450 knots, and an altitude of 300 feet.

*'**volvo**'* is considered as a vehicle interior noise, which was recorded in Volvo 340. It was moving in 4th gear, on an asphalt road, in rainy conditions.

'***factory***' was recorded in a car production hall and is considered as a factory noise. The simulation results are shown in Fig.6 to Fig.9.

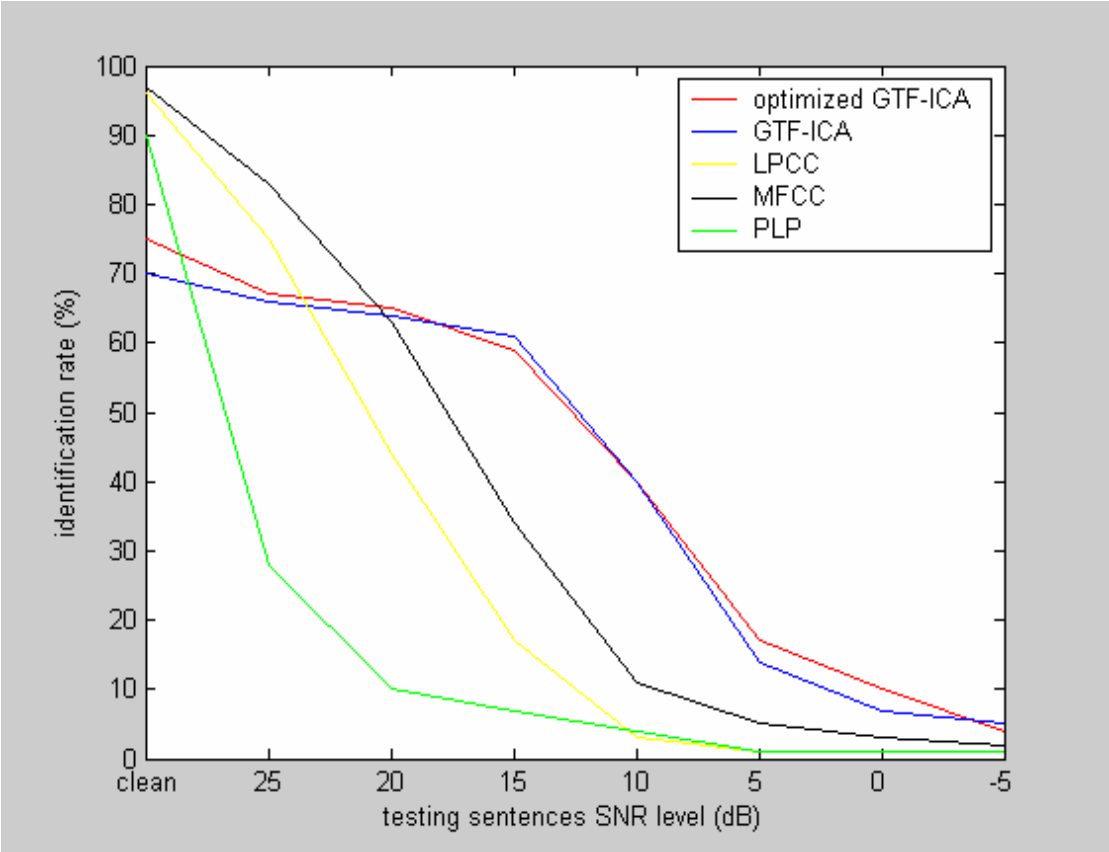

Figure 5 the resulted identification rate for clean training and testing corrupted by additive Gaussian white noise (%)

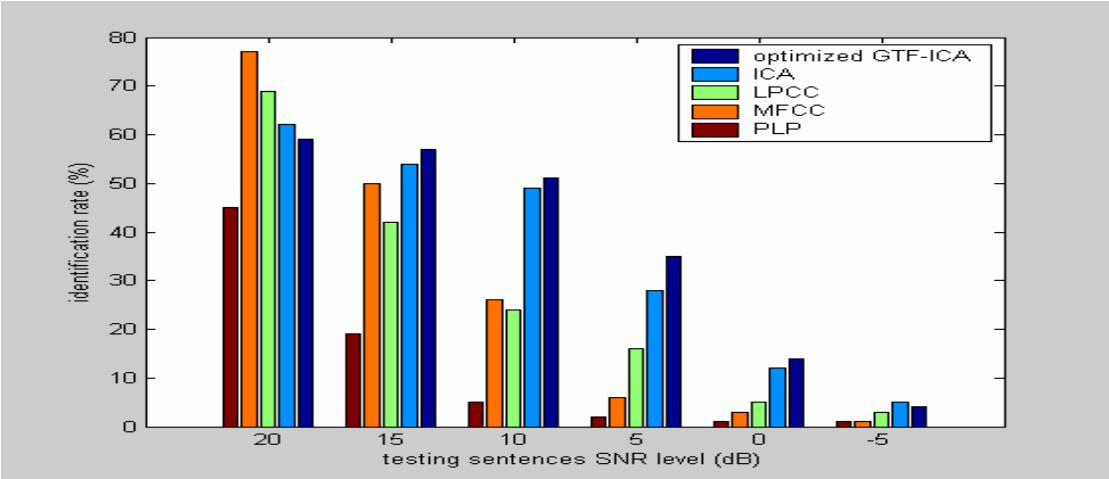

Figure 6 the resulted identification rate for clean training and testing corrupted by additive 'detoryerops' noise (%)

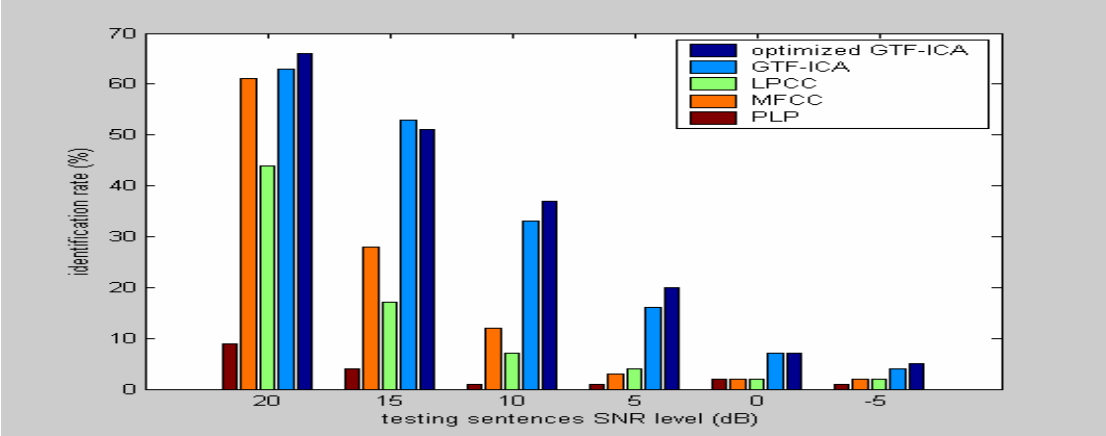

Figure 7 the resulted identification rate for clean training and testing corrupted by additive ‘buccaneer’ noise (%)

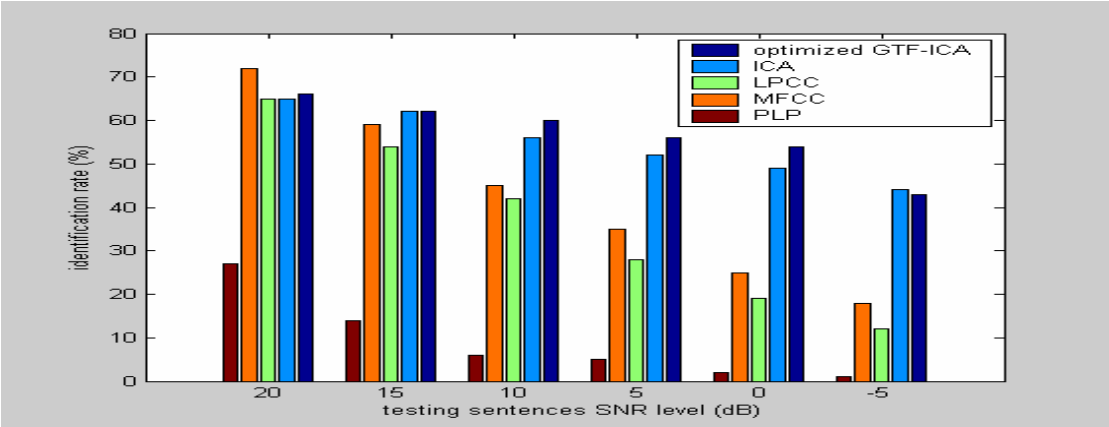

Figure 8 the resulted identification rate for clean training and testing corrupted by additive ‘volvo’ noise (%)

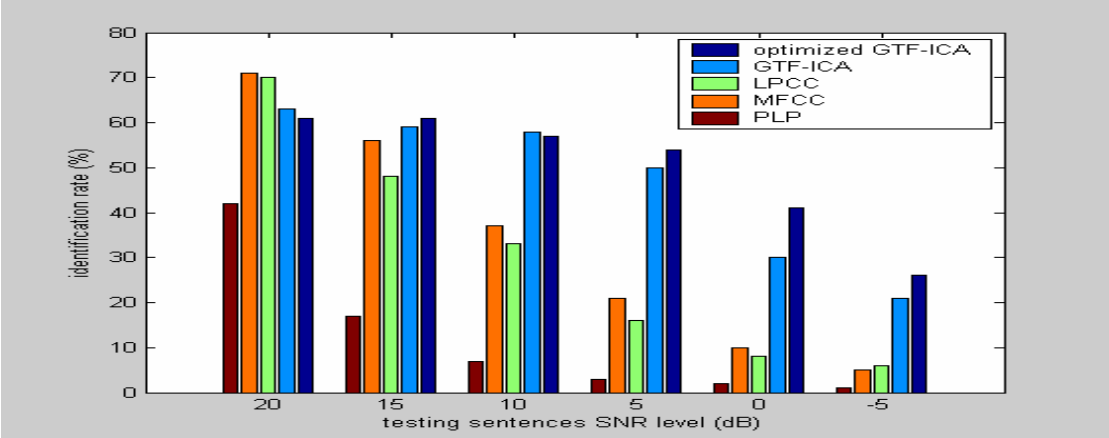

Figure 9 the resulted identification rate for clean training and testing corrupted by additive ‘factory’ noise (%)

Some observations can be obtained from the experiments:

MFCC and LPCC have the best performance while environments of both training and testing are clean, followed by PLP, and GTF-ICA and its optimized method seem to have the worst performance. However when the speaker identification system moves to the situation where the training environment is clean and the testing environment is noisy, performances of LPCC, MFCC and PLP degrades rapidly, especially for PLP, which proves that features extracted using LPCC, MFCC and PLP are significantly affected by the environment mismatch and environmental additive noises. Compared with them, GTF-ICA and its optimized method are more reliable and robust for environment mismatch. As the testing SNR level exceed 20dB, GTF-ICA feature matrix shows its advantage against other commonly used techniques and holds the better performance. At the same time, the optimized GTF-ICA method has the best performance, which proves that combining the GTF-ICA feature matrix with its dynamic delta matrix do improve the accuracy of the representation of a individual speaker.

Another advantage of using GTF-ICA feature matrix for speaker identification system is that its testing computation time is much shorter than that the other commonly used feature extraction techniques have. As mentioned before, our new method use whole testing sentence to extract the feature matrix. However for LPCC, MFCC and PLP, the testing sentence has to be segmented and extract the feature vector from each segment. On the other hand, the feature matrix extracted using GTF-ICA can directly denotes the distribution of speaker. On the contrary, GMM must be generated according the feature vectors extracted from traditional techniques to represent the speaker. Finally, for pattern classification, the maximum log-likelihood estimation used in traditional techniques is much complex than correlation calculation used in our new method. Thus the consumed time for identifying the target speaker is reduced significantly, which meet the requirement of modern speech processing application: fast and efficient.

6. CONCLUSIONS

A new technique based on Gammatone auditory filterbank and Independent Component Analysis is developed. The GTF-ICA feature matrix extracted using this technique denotes the statistical structure of speech signal in different frequency bands, and these frequency bands match the resolution of human hearing. The GTF-ICA feature matrix was tested in a speaker text-independent identification system, compared with some commonly used feature extraction techniques: LPCC, MFCC and PLP, results prove that it is more robust to mismatch environment with low SNR level. At the same time, using new method, the consumed calculation time of testing phase decreases significantly.

REFERENCES

[1] R. L. Klevans and R. D. Rodman, “Voice Recognition”, Artech House, Boston.London, 1997.

[2] H. Gish and M. Schmidt, “Text-independent Speaker Identification”, Signal Processing Magazine, IEEE, Vol. 11, Issue: 4, pp: 18-32, Oct. 1994.

[3] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis for Speech”, J. Acoust. Soc. Am., pp. 1738-1752, 1990.

[4] W. H. Abdulla, “Auditory Based Feature Vectors for Speech Recognition Systems”, In Advance in Communication and Software Technologies, N.E. Mastorakis & V.V. Kluev, Editor, WSEAS Press, pp 231-236,2002.

[5] T. W. Lee and G. J. Jang, “Learning Statistically Efficient Features for Speaker Recogntion”, Acoustics, Speech, and Signal Processing, IEEE International Conference, Vol. 1, pp:105-108, May, 2001.

[6] J. H. Lee, H. Y. Jung, T. W. Lee and S. Y. Lee, “Speech Feature Extraction using Independent Component Analysis”, Acoustics, Speech and Signal Processing, IEEE International Conference, Vol. 3, pp:1631-1634, June, 2000.

[7] J. W. Picone, “Signal Modeling Techniques in Speech Recognition”, Proceedings of the IEEE, Vol. 81, Issue: 9, pp: 1215-1247, Sept. 1993.

[8] R. Lawrence and J. Biing-Hwang, “Fundamentals of Speech Recognition”, Prentice Hall PTR, Upper Saddle River, New Jersey, 1993.

[9] A. Hyvärinen, J. Karhunen and E. Oja, “Independent Component Analysis”, JOHN WILWY & SONS, ING, 2001.

[10] J. V. Stone, “Independent Component Analysis: A Tutorial Introduction”, The MIT Press, Cambridge, Massachusetts, London, England, 2004.

[11] D. A. Reynolds and R. C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, Speech and Audio Processing, IEEE Transaction on, Vol. 3, Issue: 1, pp:72-83, Jan. 1995.

[12] H. M. Bae, H. Y. Jung, T. W. Lee and S. Y. Lee, “Subband-based Blind Signal Separation for Noisy Speech Recognition”, Electronic Letters, IEEE, Vol. 35, Issue 23, pp: 2011-2012, Nov, 1999.

[13] J. H. Lee, H. Y. J, T. W. Lee and S. Y. Lee, “Speech Enhancement with MAP estimation and ICA-based Speech features”, Electronic Letters, IEEE, Vol. 36, Issue 17, pp: 1506-1507, Aug, 2000.

[14] H. M. Park, H. Y. Jung, S. Y. Lee and T. W. Lee, “On Subband-base Blind Separation for Noisy Speech Recognition”, Neural Information Processing, IEEE, Vol. 1, pp:204-209, Nov, 1999.

[15] T. W. Lee and G. J. Jang, “ The Statistical Structures of Male and Female Speech Signals”, Acoustic, Speech and Signal Processing, IEEE International Conference, Vol. 1, pp:105-108, May, 2001

[16] L. Potamitis, N. Fakotakis and G. Kokkinakis, “Independent Component Analysis Applied to Feature Extraction for Robust Automatic Speech recognition”, Electronic Letters, IEEE, Vol. 36, Issue 29, pp:1977-1978, Nov, 2000.

[17] A. Dabrowski, D. Cetnarowicz, and T. Marciniak, “Analysis of Speech Separation for ASR systems”, Robot Motion and Control, Proceedings of the Fourth International Workshop, pp:345-350, June, 2004.

[18] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, “Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals”, Speech and Audio Processing, IEEE Transactions on vol. 11, Issue 3, pp: 204-215, May, 2003.

[19] J. H. Zhao, J. M. Kuang and X. Xie, “Data-driven Temporal Processing using Independent Component Analysis for Robust Speech Recognition”, Signal Processing and Information Technology, Proceedings of the 3rd IEEE International Symposium, pp: 729-732, Dec, 2003.

[20] J. H. Lee, T. W. Lee, H. Y. Jung and S. Y. Lee, “On the Efficient Speech Extraction Based on Independent Component Analysis”, Kluwer Academic Publisher, Vol. 15, Issue 3, pp: 235-245, Jun, 2002.

[21] O. W. Kwon and T. W. Lee, “Phoneme Recognition using ICA-based Feature Extraction and Transformation”, Elsevier North-Holland, Inc, Vol. 84, Issue 6, pp:1005-1019, Mar, 2004.