Voice biometric feature using Gammatone filterbank and ICA

Waleed H. Abdulla* and Yushi Zhang

Department of Electrical and Computer Engineering,
The University of Auckland,
Private Bag 92019,
Auckland, New Zealand
E-mail: w.abdulla@auckland.ac.nz
E-mail: yzha316@hotmail.com
*Corresponding author

Abstract: Voice biometric feature extraction is the core task in developing any speaker identification system. This paper proposes a robust feature extraction technique for the purpose of speaker identification. The technique is based on processing monaural speech signal using human auditory system based Gammatone Filterbank (GTF) and Independent Component Analysis (ICA). The measures used to assess the robustness to additive noises and speaker identification performance are defined and discussed. The kkn the proposed feature is evaluated in real environments under varying noisy conditions. The proposed feature is benchmarked against the commonly used features such as: MFCC, PLCC, and PLP, and it outperforms them in different noisy environments.

Keywords: human biometrics; speaker modelling; speaker recognition; voice recognition; voice biometrics; Gammatone filterbank; independent component analysis; speech feature extraction; processing noisy speech.


Biographical notes: Waleed H. Abdulla holds a PhD from the University of Otago, New Zealand. Since 2002, he has been working in the Department of Electrical and Computer Engineering/The University of Auckland. His main research interests are in human biometrics, speech and audio signal processing. He has been serving as Editorial Board member of several journals and Vice-President of Member Relations and Development of APSIPA Association. He published more than 80 refereed publications including a patent and a book, and supervised more than 25 postgraduate students. He has been recently awarded JSPS. He received the award for Excellent Teaching for the year 2005. He is a member of ISCA, IEEE and IET.

Yushi Zhang received his PhD Degree from The University of Auckland in 2009. His main interest is in Speaker identification systems and he has many publications in this line of research.
1 Introduction

Voice biometric carries information that highly characterises speakers. This information comprises identity, gender, accent, age group, mood, and more. In this paper, our main concern is to extract robust voice biometric feature to implement speaker identification systems. Speaker identification is one of the highly demanded systems the market has been looking for since decades. It generally constitutes the core part of many security systems. Security applications may include automatic identification to avoid: Intruders lurking into buildings; Carjackers walk away with cars; Unauthorised banking access; Hackers accessing computers and many more.

Voice biometric attains attraction because it can be embodied in applications based on normal landline or mobile phones, such as banking systems. The word ‘biometric’ is derived from the Greek words ‘bios’ and ‘metric’, which mean life and measurement, respectively. This directly translates into: ‘life measurement’. Voice biometrics is partially physiological as it depends on the vocal tract shape and the vocal cord structure. However, it is mostly behavioural as speech manifested through the vocal tract dynamics and the way people talk.

Voice biometric is attributed with major advantages over the other types of identification techniques in terms of practicality, smooth deployment, user acceptance, and is highly non-intrusive as compared to the other non-speech biometrics.

The spectral based features have many advantages that make them widely used in speaker recognition systems. The main advantage of spectral methods is that logarithmic scales can apply to either amplitude or frequency to imitate the function of the peripheral auditory process of the human ear and improve the recognition rate.

The information about the anatomical structure of the vocal apparatus is widely used as speech feature for speaker recognition as it is easy to extract in an automatic fashion. The idea is taken from the fact that each speaker has his or her unique vocal structure, similar to the retina or fingerprint. This information can be represented by the spectra of a speech signal since different speakers will have different spectra (location and magnitude of peaks) for similar sounds. Spectral feature extraction is usually computed by common methods like Linear Predictive Coding (LPC) (Klevans and Rodman, 1997), Log Area Ratio (LAR) (Lawrence and Biing-Hwang, 1993), Linear Predictive Cepstral Coefficients (LPCC) (Lawrence and Biing-Hwang, 1993), Mel Frequency Cepstral Coefficient (MFCC) (Gish and Schmidt, 1994), Perceptual Linear Predictive (PLP) (Hermansky, 1990), and Relative Spectra (RASTA) (Hermansky and Morgan, 1994). Among all these features, the most popular techniques are LPCC, MFCC, and PLP. Thus, we will use them for benchmarking the proposed feature in this paper.

In this paper we propose a new spectral technique to extract robust-to-noise feature which acts as a front end for speaker identification systems. It is based on Gammatone filterbank and independent component analysis in the frequency domain (F-GTF-ICA), where ‘F’ refers to the frequency domain.

This paper is structured as follows. Section 2 introduces the proposed feature for speaker identification based on voice biometrics. Section 3 is for evaluating the sensitivity of the feature to different kind of noises. Section 4 details the classification method for speaker identification as well as the measures considered to reduce the computational cost. Section 5 is for benchmarking the proposed feature against the commonly used features. Section 6 is for evaluating the performance of the system under real environmental noises. Section 7 derives conclusions from the conducted research.
2  F-GTF-ICA feature extraction technique

The F-GTF-ICA technique is an attempt to imitate the human auditory system in some aspects. Its main structure comprises Gammatone Filterbank, rectification, compression, and ICA in the frequency domain. The Gammatone Filterbank emulates the human cochlea frequency resolution. The rectification and compression stages are to emulate the behaviour of the inner hair cells, while the ICA stage is to extract the dominant frequency components from the filterbank. The extracted feature emphasises the differences in the statistical structures among the speakers, and can model the distribution of the individuals. The block diagram of the proposed front end is illustrated in Figure 1 (Zhang and Abdulla, 2006).

Figure 1  F-GTF-ICA feature extraction (see online version for colours)

2.1  Gammatone filterbank structure

The speech signal is first segmented into 30 ms frames and multiplied by a hamming window with 50% overlap then the output is forwarded to a Gammatone filterbank. The Gammatone filter offers a means of producing an auditory filterbank that conforms well to both physiological and psychophysical data on frequency selectivity (Cooke, 1993). It models the function of the human cochlea by a bank of overlapping bandpass filters. The impulse response of each filter follows the Gammatone function shape and has the following classical form (Aertsen and Johannesma, 1980):

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

(1)

where, \( \gamma(n,b) \) is a normalisation constant depending on the filter order, \( n \), and the bandwidth related factor, \( b \), \( \omega \) is the radian centre frequency, \( \phi \) is the phase shift and \( u(t) \) is a unit step function to achieve the causality of the filter. For the sake of computation simplification, the cosine term is removed from the above equation (Cooke, 1993; Solbach, 1998). The Gammatone function corresponding to a 4th order cochlea filter centred at 1000 Hz and bandwidth of 125 Hz is shown in Figure 2.

Figure 2  The Gammatone function (see online version for colours)
The bandwidth of each filter of the Gammatone filterbank can be 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 (Abdulla, 2002). In 1938, Fletcher (Allen, 1995) firstly proposed the determination of the CB. In his algorithm, the shape of the auditory filters was assumed to be rectangular. The formulation of the signal and the noise power within the CB then is greatly simplified due to this assumption. The rectangular critical band concept is very useful, even if it is not realistic. The determination of bandwidth of the actual auditory filters can be related to it, by proposing an Equivalent Rectangular Bandwidth (ERB) filter with a unit height and a bandwidth ERB. It passes the same power as the real filter does when subjected to a white noise input. This definition of ERB implies the mathematical formula:

$$\text{ERB} = \int_0^\infty |H(f)|^2 \, df$$

(2)

where the maximum value of the filter transfer function, $|H(f)|$, is unity. Several psychoacoustically motivated formulae have been derived for the ERB values. In this paper, our preference is for that suggested by Glasberg and Moore (1990) and corrected by a factor 1.019 by Patterson (1994). It follows the following formulae at centre frequency $f_c$:

$$\text{ERB} = 1.019 \times 24.7 \left( 1 + \frac{4.37}{1000} f_c \right).$$

(3)

This formula gives the highest selectivity factor, $Q$ factor (defined as the ratio between the centre frequency and the bandwidth of the each filter), among all the other suggested ones (Abdulla, 2002). Thus, to determine the bandwidth of each filter, which is now represented by the ERB value, we have to determine the centre frequency of each filter beforehand. Along the 35 mm spiral path cochlea of the human ear, there are around 3000 inner hair cells. Each one 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 is impractical to implement. However, our objective can be achieved by specifying a certain number of overlapping filters. The percentage-overlapping factor, $v$, specifies the number of channels, filters, required to cover the useful frequency band (Abdulla, 2002). If we consider that the information carrying band is bounded by $f_H$ Hz and $f_L$ Hz, then the number of filters will be:

$$N = \frac{9.26}{v} \ln \frac{f_H + 228.7}{f_L + 228.7}.$$  

(4)

Then the centre frequency can be calculated by

$$f_c = -228.7 + (f_H + 228.7)e^{-vn/9.26}$$

(5)

where $1 \leq n \leq N$.

Having decided the location of the centre frequency of each filter, the bandwidth can be calculated from equation (3) and we can now proceed to the implementation stage. The implantation of a bandpass filter from its time domain function is a straightforward procedure in signal processing (Abdulla, 2002). It is simply started by finding the Laplace transform of the Gammatone function, then mapping it into the digital form by using pole-mapping, bilinear transform or impulse invariant transform. There are several
methods for representing and implementing the Gammatone function. Lyon (1997) suggested an all-pole version which discarded the zeros from the transfer function of the Gammatone filter and had the advantage of simple parameterisation. His all-pole version reduces the computation, but as a resulting trade-off, selectivity sharpness of low frequency is lost. Another form was suggested by Cooke (1993), in which a fourth-order filter is realised by using the complex domain. His method needs pre-multiplication of the input signal by a complex exponential at the specified centre frequency, filtering with a base-band Gammatone filter, post-multiplication by that exponential. Slaney (1993) described one simple way of implementation procedure of the Gammatone based filters. In this implementation, a pole-mapping technique is used to convert from the continuous domain Gammatone response to the digital domain. In our approach, our preference is for the Slaney method, because it preserves the original form of the Gammatone filter and also the simplicity of implementation.

The frequency response of a 30-channel filterbank, covering the 200–11,025 Hz band, is shown in Figure 3. The bandwidth of the channel is logarithmically proportional to 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.

Figure 3 Frequency response of a 30-channel filterbank covering 200–11,025 Hz (see online version for colours)

2.2 Rectification and compression

Now, assume that the number of Gammatone filters is $N$. Thus, the output of the Gammatone filterbank is a matrix $Y$, which has $N$ rows, and each row represents the output of each bandpass filter of Gammatone filterbank in time domain and has the same sampling length as the analysis frame does. Each bandpass filtered version of the original signal is followed by a half-wave rectifier and a power-law compressor, simulating the behaviour of inner hair cells. The task of the inner hair cells is the so-called transduction process, that is, to convert mechanical movements into electrical potentials. It is assumed that the displacement of the cilia of the cells is proportional to the basilar membrane velocity. Measurements of electrical responses have revealed a directional sensitivity; while displacement in one direction is excitatory, movement in the opposite direction is inhibitory (Dallos et al., 1996). Thus, the cells mainly react to give a positive deflection of the basilar membrane and, consequently, it is reasonable to model this behaviour with
a half-wave rectifier. Therefore, half-wave rectification is commonly used to model this aspect of physiology (Lyon, 1982; Seneff, 1988).

The aforementioned measurements also show a compressive response (Dallos et al., 1996). Therefore, we apply a power-law compressor to the half-wave rectified signals. Two particular forms of compressions are used in practice; the μ-law and A-law. For a given signal x, the output of the μ-law compressor is:

\[ y(\mu) = \frac{V \log(1 + \mu|x|/V)}{\log(1 + \mu)} \ \text{sgn}(x). \]  

(6)

While the output of the A-law compressor is:

\[ y(A) = \begin{cases} \frac{A|x|}{1 + \log A} \ \text{sgn}(x) & 0 \leq |x| \leq \frac{V}{A} \\ \frac{V(1 + \log(A|x|/V))}{1 + \log A} \ \text{sgn}(x) & \frac{V}{A} < |x| \leq V \end{cases} \]  

(7)

where μ and A are the μ-law and A-law parameters, respectively, and V is the maximum value of the signal x. In Figure 4, the μ-law and A-law for different parameters are shown. We see, therefore, that both of the power laws are neither strictly linear nor strictly logarithmic, but they are approximately linear for low value inputs and approximately logarithmic for high value inputs (Haykin, 2000).

**Figure 4** Compression laws: (a) μ-law and (b) A-law (see online version for colours)

For both the μ-law and A-law, the dynamic range capability of the compressor can be improved by increasing the parameters μ and A, respectively. However, the SNR for low-level signals increases at the expense of the SNR for high-level signals (Haykin, 2000). To accommodate these two conflicting requirements, a compromise is usually made by carefully choosing particular values for μ and A. The typical values used in practice and also in our research are: μ = 255 and A = 87.6 (Haykin, 2000). The A-law and μ-law compressors result in almost similar identification performances because these
two power laws have nearly the same compression characteristics. In our research, a half-wave rectifier followed by the A-law compressor is adopted.

2.3 ICA basis application

The Fourier transform of the rectified and compressed filterbank outputs is taken to imitate the human aural discrimination capability which manifests itself in the frequency domain. Afterward, the absolute value of the frequency spectra is performed to introduce $X$, indicated in Figure 1. The phase information, and, likewise, the human auditory system, are discarded as they have no significant effect on the speaker identification system performance (Chow, 2004; Sanderson, 2002).

To extract the independent components of speech signals, the ICA algorithm is applied to the observation $X$ as shown in Figure 1 (Hyvärinen et al., 2001). Here, $X$ represents a spectral matrix and each row $x_i$ represents the absolute Fourier spectra of the $i$th rectified and compressed output of the Gammatone filterbank.

$$X = A \cdot S = \sum_{i=1}^{N} a_i s_i$$  \hspace{1cm} (8)  

$$S = A^{-1} \cdot X = W \cdot X = \sum_{i=1}^{N} w_i x_i$$  \hspace{1cm} (9)

where $A$ and $W$ are $N \times N$ scalar square matrices and they denote the mixing and unmixing matrices respectively, the column vectors $a_i$ and $w_i$ are basis functions generating the observed signal $X$ and components $S$ respectively, $N$ is the number of filterbank channels.

The extracted basis functions $w_i$ capture the correlations among the frequency components of the speech signal. This is achieved by ICA in the form of linear combinations of basis filter functions specific to each person. Those correlations can be considered as functions of a speaker’s glottal or nasal shape (Holmgren, 1967). Therefore the F-GTF-ICA feature matrix $W$ is specific to individuals. Meanwhile, ICA leads a highly efficient representation of the speech signal. It does not only decorrelate the second order statistics but also reduces the higher-order statistical dependencies. Hence it captures the main and essential underlying factors from statistical data. These factors represent the statistical structure of the data and are slowly varying among the data (Hyvärinen et al., 2001). Considering a speech signal, the slow varying components may present the information about gender, accent, age, speaking rate, and phonetics realisations of the speaker. This speaker related information can be assumed to be pseudo stationary information of the speech signal. On the other hand, the fast varying components may reflect some linguistic information which varies all the time while speaking. Thereby, ICA may extract more speaker related information than linguistic information. As a result, we may argue that the F-GTF-ICA feature matrix embodies rich speaker’s attributes. Furthermore, the feature matrix represents the statistical structure of the speech signal in different frequency bands and these frequency bands are taken from the Gammatone filterbank which is designed to imitate the frequency resolution of human hearing. Hence, this new feature technique also introduces concepts of the human aural system to the processing, which humanises the speaker identification system and makes the system more reliable.
3 F-GTF-ICA feature matrix evaluations

It is well known that speech signal contains not only linguistic information but, also, speaker dependent information. At the same time, it may contain noisy information when the speech signal is contaminated with noise. Therefore, any feature extracted from the speech signal characterises both linguistic information as well as speaker attributes, and it is distorted by additive noises. In the case of text-independent speaker identification, our objective is to design features to fill the feature space that depends primarily on speakers; not the particular text spoken, and robust to additive noises. Intensive simulations are carried out to verify this objective.

3.1 Performance evaluation of the proposed feature

The performance of the F-GTF-ICA feature matrix is investigated in this section. Two speakers’ speech signals are randomly selected from the TIMIT database. The TIMIT speech database was recorded by a high quality microphone in a noise free laboratory environment. The bandwidth of the speech signals is 8 KHz and the average signal to noise ratio (SNR) is about 53 dB. TIMIT database contains utterances of 630 speakers from eight different dialects of spoken English, and for each speaker there are a total of ten sentences arranged in three categories (dialectic calibration, random contextual variant and phonetically compact sentences). The average duration of each sentence is about 3 s. For each speaker, all the speech signals were recorded in a single session. TIMIT is used to test the text-independent speaker identification performances in an ideal situation. There is no noise, no channel variation and no intersession variation. The F-GTF-ICA feature matrices are extracted from 30 ms frames segmented from the speakers’ utterances. A 30-channel Gammatone filterbank is adopted since that can best characterise the human aural processing for a speech signal sampled at 16 KHz.

The template feature matrix of each speaker is calculated by taking the mean of the segmented speech F-GTF-ICA feature matrices. Two situations are considered: the same speaker speaks different texts, and different speakers speak the same text. After that, the extracted feature matrices are compared by measuring their similarity. The similarity of the two matrices is quantified by calculating the cross-correlation coefficient between them. A cross correlation coefficient is a standard method of estimating the degree to which two series are correlated. The cross-correlation coefficient $\rho_{X,Y}$ between two random variables $X$ and $Y$ realised at $x_i$ and $y_i$, where $i = 1, 2, ..., N$, is formulated by:

$$
\rho_{X,Y} = \frac{\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i^2)} \sqrt{\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i^2)}}.
$$

(10)

The value of the cross-correlation coefficient is between 0 and 1, and a higher value means a higher similarity. The final value of the cross-correlation is the average value of the cross-correlation from each column. Table 1 summarises the resulting cross-correlation coefficient and the correlation ratio (ratio between the cross-correlation coefficient produced by a speaker with different texts and that produced by different speakers speaking the same text). For comparison, the MFCC feature is also included in the table. Apparently, even when the same speaker speaks different texts, the F-GTF-ICA feature matrices produced from these two utterances introduce a higher degree of
similarity than that produced by different speakers speaking the same text. This proves that the F-GTF-ICA feature matrix contains more speaker related information than linguistic information, which is a desirable property for speaker identification systems. Meanwhile we find that the MFCC features produce a high degree of similarity when the same speaker speaks different utterances, which is similar to that produced by F-GTF-ICA. However, MFCC also produces high degree of similarity when different speaker speak the same text. That shows MFCC contains close levels of speaker related and linguistic information at the same time, which degrades the speaker identification performance.

<table>
<thead>
<tr>
<th>Utterance duration</th>
<th>Same speaker/different texts</th>
<th>Same text/different speakers</th>
<th>Correlation ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-GTF-ICA</td>
<td>MFCC</td>
<td>F-GTF-ICA</td>
</tr>
<tr>
<td>30 milliseconds</td>
<td>0.6325</td>
<td>0.4534</td>
<td>0.4200</td>
</tr>
<tr>
<td>1 second</td>
<td>0.6940</td>
<td>0.5398</td>
<td>0.4891</td>
</tr>
<tr>
<td>2 seconds</td>
<td>0.7581</td>
<td>0.7489</td>
<td>0.5147</td>
</tr>
<tr>
<td>3 seconds</td>
<td>0.8396</td>
<td>0.8542</td>
<td>0.5254</td>
</tr>
<tr>
<td>4 seconds</td>
<td>0.8530</td>
<td>0.8723</td>
<td>0.5362</td>
</tr>
<tr>
<td>5 seconds</td>
<td>0.8695</td>
<td>0.8834</td>
<td>0.5840</td>
</tr>
<tr>
<td>6 seconds</td>
<td>0.8763</td>
<td>0.8932</td>
<td>0.6432</td>
</tr>
</tbody>
</table>

3.2 Robustness evaluation of the proposed feature

We compare the robustness of feature matrix with other commonly used feature extraction techniques, such as LPCC, MFCC and PLP. The simulation is carried out as follows: reference feature vectors (or matrices) are generated from a 5-second clean speech signal of a speaker selected from the TIMIT database; afterwards, various noises are added to this speech signal to produce noisy speech and, accordingly, noisy feature vectors (or matrices). Then we compare them with the reference (clean) feature to substantiate the distortion caused by the additive noises. Therefore, we must have a quantitative measure of this feature distortion. Here we introduce the similarity degradation measure. Again, the similarity between two feature vectors (or matrices) is measured by calculating the cross-correlation coefficient between them. Thus, the feature distortion is defined by the following equation:

\[
\text{feature distortion} = 1 - \rho_{\text{clean, noisy}} 
\]

where, \( \rho_{\text{clean, noisy}} \) is the value (between 0 and 1) of the cross-correlation coefficient between clean and noisy feature vectors (or matrices). Therefore, the feature distortion also varies between 0 and 1. A high value denotes a high distortion. The additive noises are white Gaussian noise and some other colour noises from NOISEX-92 database i.e., factory, vehicle interior, destroyer operations background, and babble noises. The NOISEX-92 is a noise database which provides various noises recorded in real environments. Factory noise is recorded near plate-cutting and electrical welding equipment. Vehicle interior noise is recorded in a moving vehicle (at 120 km/h, 4th gear)
on an asphalt road and in rainy conditions. Destroyer operations background noise is recorded while a destroyer machine was working. The source of babble noise is 100 people speaking in a canteen. The canteen radius is over two metres; therefore, individual voices are slightly audible. The simulation results under the effect of these noises are summarised in Figure 5. It is evident that the F-GTF-ICA feature has less distortion compared with LPCC, MFCC, and PLP features when the speech signal is contaminated by additive noises, except in the situation where testing utterance includes babble noise with SNR above 14 dB. Therefore, in most cases, the F-GTF-ICA feature is less sensitive to additive noises. A possible explanation why the proposed feature does not work well in a babble noisy environment may be that F-GTF-ICA extracts speakers’ related information from all speech like signals and thus develops distorted feature matrices.

**Figure 5** Feature distortion caused by different kinds of noises (see online version for colours)
4 Diagonal F-GTF-ICA feature matrix

From Section 1, it has been realised that we can extract one F-GTF-ICA feature matrix from each speech frame. The template feature matrix specific to a certain speaker is then found by taking the mean of these frame based feature matrices. The template feature matrix is then used for classification in the speaker identification system. We use a simple algorithm for patterns classification. The idea of the algorithm is that the basis functions matrix $W$ is estimated from the observation matrix $X$ such that the random variables $s_i$ are as independent as possible.

During the training process, assume that the template feature matrix $W^\text{train}$ is developed from a speech data of a certain speaker. Intuitively, test data from the same speaker will show low degree of independency when the test data are projected onto this trained feature matrix. However, if test data from a different speaker are used, the degree of independency will be higher (Rosca and Kofmehl, 2003). The feature matrix is not designed to minimise the independency on data coming from a speaker characterised by a different correlation structure in the frequency domain. Therefore, we can define the identification score $\bar{\eta}$ by:

$$S^\text{test} = W^\text{train} \cdot X^\text{test}$$  \hfill (12)

$$\Gamma_{\eta^\text{max}}(S^\text{test}) = \sum_{i=1}^{N} \left| r_{ij} \right|^\beta$$  \hfill (13)

where $r_{ij}$ is the cross-correlation coefficient between the random variables $s_i$ and $s_j$, which reflects the degree of dependency between them. $\beta$ is a positive constant and chosen empirically to be 2.

For speaker identification, a group of $M$ speakers is represented by their F-GTF-ICA template feature matrices $W_1^\text{train}, W_2^\text{train}, \ldots, W_M^\text{train}$. The identity of the test speaker is determined by the minimum value of $\Gamma$:

$$D_k = \arg \min_k (\Gamma_{\eta^\text{max}}(S^\text{test})) \quad 1 \leq k \leq M$$  \hfill (14)

where $D_k$ can be treated as the distance measure between the test and the $k$th training matrix, $k$ is the index of the training speaker.

However, this method has a large computational cost. The computational bottleneck is in the calculation of the cross-correlation coefficients between the random variables $s_i$ and $s_j$ to check their independency. For a feature matrix of dimension $N \times N$, there are $(N + (N - 1) + (N - 2) + \ldots + 1) = N/2(N + 1)$ cross-correlation coefficients to be calculated. One way to reduce the computational cost is to discard the ineffective components in the feature matrix.

Figure 6 shows the distribution of the F-GTF-ICA feature matrix elements extracted from a speech frame. Clearly, the most noticeable elements are located upon the diagonal area of the matrix. This led us to consider some of the diagonal elements for calculations while discarding all the off-diagonal elements.
Voice biometric feature using Gammatone filterbank and ICA

Figure 6 3-D surface of an F-GTF-ICA feature matrix distribution (see online version for colours)

Figure 7 shows the elements of the example F-GTF-ICA feature matrix. In the adopted approach, we add the cross diagonal elements between the two dashed red lines, that border the effective elements, located on the blue solid line directions. This would introduce a diagonal matrix whose diagonal elements are the summation of a certain number of the cross diagonal elements. This later matrix is to be used for further modelling and pattern classification instead of the full elements matrix. Now the question is: How many cross diagonal elements in the feature matrix should be used? The $F$-ratio is used to decide this factor. $F$-ratio is a measure that can evaluate the effectiveness of the feature coefficients. It has been widely used as a figure of merit for feature selection in speaker recognition applications (Chow, 2004). In the contest of feature selection for pattern classification, the $F$-ratio can be considered as a strong catalyst to select the features that maximise the separation between different classes. It can be formulated by:

$$F\text{-ratio} = \frac{\text{between-class variance}}{\text{within-class variance}}$$

In our experiment, 188 F-GTF-ICA feature matrices produced from 188 different speakers from the TIMIT database are used to calculate the F-ratio values for different diagonal areas. Figure 8 illustrates the $F$-ratio values of the diagonalised F-GTF-ICA feature matrix.

Figure 7 Diagram of an F-GTF-ICA feature matrix (see online version for colours)
From Figure 8, the highest $F$-ratio is obtained when five cross diagonal elements are added to form one diagonal element, which indicated that this developed feature matrix contains rich speaker information. Near the upper and lower corners of the matrix, there are less than five cross diagonal elements and we only use those available elements. The $F$-ratio decreases when more elements are included. That is because the additional elements may contain some speaker irrelevant information, which may be speech content information, or noisy information. Therefore, in our proposed method, the sum of five cross diagonal elements from F-GTF-ICA feature matrix is used. Then the resulted feature vector is used as inputs to build Gaussian Mixture models (GMM) which will be used for pattern classification to identify the unknown speaker. Thus, the computation load is substantially reduced by using the maximum likelihood estimation using the diagonal matrices instead of calculating a large number of cross-correlation coefficients from the full matrices.

5 Performance evaluations

In this section, text-independent speaker identification tasks are carried out to evaluate the performance of the proposed algorithm. In our proposed system, 188 speakers (100 males and 88 females from eight different dialects) from TIMIT are used. The length of the training sentences is about 15 s while that of the testing sentences is 5 s. The speech signal is first segmented into 30 ms frames multiplied by a hamming window with 50% overlap, and a 30-channel F-GTF-ICA feature matrix is extracted using our proposed method from each frame. After that, five cross diagonal elements of each feature matrix are added to form the feature vectors (or diagonal matrices). In our speaker identification system, the feature vectors are used to form 32-component Gaussian Mixture Models (GMM) to classify individuals. To compare the proposed F-GTF-ICA algorithm with the other commonly used features, we generated a baseline system using 24 coefficients MFCC, LPCC and PLP feature vectors (without using the delta coefficients). The identification rate is defined by:
\[
ID \text{ rate} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100
\]

where \(N_{\text{correct}}\) denotes the number of the correctly identified testing speakers and \(N_{\text{total}}\) is the total number of training speakers.

5.1 Performance evaluation in clean environment

In this experiment, we test our proposed algorithm in an ideal situation: both training and testing sentences are recorded in clean environments. Table 2 lists the identification results of different feature extraction techniques.

<table>
<thead>
<tr>
<th>Feature</th>
<th>F-GTF-ICA</th>
<th>LPCC</th>
<th>MFCC</th>
<th>PLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID rate (%)</td>
<td>98.94</td>
<td>96.81</td>
<td>98.94</td>
<td>91.49</td>
</tr>
</tbody>
</table>

This table shows that the best result is obtained by F-GTF-ICA and MFCC methods. Interestingly, it was noticed that the same two speakers are misidentified by using both of these two methods. This proves that the GTF-ICA feature matrix performs similar to the MFCC, and it efficiently represents the variability of speakers and denotes the distribution of individuals.

5.2 Performance evaluation in a noisy-mismatched environment

In this experiment, the effect of a noisy-mismatched environment is investigated. Here, we refer the noisy-mismatched environment to the situation where clean utterances are used for training, while noise-additive utterances are used for testing. The additive noises are brought from the NOISEX-92 database which includes many types of noises such as Gaussian, colour, factory, vehicle interior, destroyer operations, babble, and more noises. In these experiments the noisy speech signal is simply prepared by adding the noise signal to the clean speech signal subjected to a certain SNR level. This is an ideal preparation of the testing noisy signal where there is no convolution noise or channel effect. We label adding noise to the speech signal in this way as studio recording condition.

The simulation results are shown in Figure 9. Apparently, for all the methods, the system performance decreases rapidly while the additive noise level increases. During these, apart from the babble noise (SNR level is around 20 dB), the F-GTF-ICA method has better performance in comparison to the other commonly used feature extraction methods, such as LPCC, MFCC and PLP, in noisy-mismatched environments. From the experiments, the F-GTF-ICA method shows its advantage of robustness when the additive noise level increases, since its identification rate reduces less than the other algorithms do.
6 Feature evaluations in real environments

The proposed feature has been proven to be an effective one, capable of achieving high identification accuracy in noisy-mismatched environments. However, there are many challenges and problems that should be addressed in applying our proposed methods to ‘real-world’ applications where additional sources of variability might exist beyond those sources of variability in the studio recording. The first step toward addressing these problems is to log speech utterances recorded in real environments. A speech corpus based on the TIMIT database has been developed. In this scenario, a source loudspeaker playbacks speech signals reproduced from the TIMIT database. Different interference and background noises with various SNR levels are generated using distributed noise loudspeakers. The mixing noises implemented are factory and destroyer operations noises taken from the Noisex-92 database. The mixed noisy speech signal was recorded via an omnidirectional electret microphone. The source loudspeaker, noise loudspeakers, and the microphone were positioned to be at the same height. The source loudspeaker stood about 30 cm away from the microphone and two other loudspeakers, used as noise sources, were located 1.5 m and 1m away from the source loudspeaker with azimuth angles of 45° and 60°, respectively as shown in Figure 10.
The prepared speech corpus includes 188 speakers producing 12 s training and 5 s testing utterances recorded in two different rooms: a non-reverberant whisper room and a normal office room. In the non-reverberant whisper room, walls, ceiling, and floor are covered by anechoic material, which can totally absorb the speech sounds. However, in the normal office room the reverberation is quite obvious and it would affect (degrade) the recognition performance. Our proposed feature extraction algorithms are tested in a speaker identification system on our developed speech corpus.

6.1 Experimental results in non-reverberant whisper room

In this experiment, speech utterances were produced and recorded in a non-reverberant whisper room. The reverberation effect in the room is minimal apart from some reflections from the microphone and loudspeakers. Two loudspeakers were used to produce the factory and destroyer background noises, respectively. The training utterances were recorded in a noise free environment while the testing speech was recorded in clean and noisy environments with various SNR levels. The experimental results if the identification rates (ID rate) are summarised in Tables 3–6.

Table 3   ID rate evaluated in non-reverberant whisper room using different features in clean environment

<table>
<thead>
<tr>
<th>Feature</th>
<th>ID rate (%) in clean environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-GTF-ICA</td>
<td>96.27</td>
</tr>
<tr>
<td>MFCC</td>
<td>96.81</td>
</tr>
<tr>
<td>LPCC</td>
<td>94.15</td>
</tr>
<tr>
<td>PLP</td>
<td>88.83</td>
</tr>
</tbody>
</table>

Table 4   ID rate evaluated in a non-reverberant whisper room: testing with factory noise active

<table>
<thead>
<tr>
<th>Features</th>
<th>ID rate (%) with SNR levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 dB</td>
</tr>
<tr>
<td>F-GTF-ICA</td>
<td>71.81</td>
</tr>
<tr>
<td>MFCC</td>
<td>65.43</td>
</tr>
<tr>
<td>LPCC</td>
<td>60.64</td>
</tr>
<tr>
<td>PLP</td>
<td>54.26</td>
</tr>
</tbody>
</table>
Table 5  ID rate evaluated in a non-reverberant whisper room: testing with destroyer operations noise active

<table>
<thead>
<tr>
<th>Features</th>
<th>ID rate (%) with SNR levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20dB</td>
</tr>
<tr>
<td>F-GTF-ICA</td>
<td>66.49</td>
</tr>
<tr>
<td>MFCC</td>
<td>65.96</td>
</tr>
<tr>
<td>LPCC</td>
<td>58.51</td>
</tr>
<tr>
<td>PLP</td>
<td>55.85</td>
</tr>
</tbody>
</table>

Table 6  ID rate evaluated in non-reverberant whisper room: testing both factory and destroyer operation noises are active

<table>
<thead>
<tr>
<th>Features</th>
<th>ID rate (%) with SNR levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 dB</td>
</tr>
<tr>
<td>F-GTF-ICA</td>
<td>63.83</td>
</tr>
<tr>
<td>MFCC</td>
<td>61.17</td>
</tr>
<tr>
<td>LPCC</td>
<td>57.45</td>
</tr>
<tr>
<td>PLP</td>
<td>54.79</td>
</tr>
</tbody>
</table>

Remarks on the experimental results:

- Compared with the experimental results of the studio recording, the identification rate drops when the speech signal is produced via a loudspeaker for all the feature extraction algorithms, as apparent from Tables 2 and 3. The reason is that speech utterances acquired by studio recording have better quality than those acquired by loudspeaker recording because the medium between the loudspeakers and microphone fades the speech signals. Actually, the SNR for clean environment in the non-reverberant whisper room is about 48.17 dB. Therefore, the fading, due to the intermediate medium separating the sources loudspeaker and microphone, and reflection of the speech signal due to the microphone and loudspeakers are the two main factors that degrade the performance of the identification system in the whisper room. From now we will refer to the situation where the utterances are produced by loudspeaker and recorded via microphone as ‘loudspeaker recording’ and the situation in which we directly use the speech signal from the database as ‘studio recording’, since the speech data have high quality and are likely to be recorded in a studio.

- In the noisy-mismatched conditions, Gammatone auditory filterbank and independent component analysis based feature extraction method F-GTA-ICA outperforms the traditional techniques, such as LPCC, PLP and MFCC, most significantly for SNR equals 10 dB and 5 dB. This confirms that the noise effect is alleviated by using ICA algorithm, since ICA gives a highly efficient representation of the speech signal. Thus the resulting feature matrices are less affected by additive noises.
6.2 Experimental results in office room

The previous experiments are repeated in a different acoustic environment, where the training utterances are recorded in a clean and non-reverberant whisper room while the testing speech signal is recorded in a normal office room. Therefore, reverberations due to the walls, ceiling, floor and furniture in the room have negative implications for the recognition performance. There are multiple delay and scaled versions of the original testing utterance, which also participate in degrading the system performance.

The results are shown in Tables 7–9. The same observations can be noticed from the experimental results in the non-reverberant whisper room. We observed that under the same noisy-mismatched conditions, identification rate decreases about 5% on average, since reverberation hurts the system performance.

<table>
<thead>
<tr>
<th>Features</th>
<th>20 dB (ID rate)</th>
<th>10 dB (ID rate)</th>
<th>5 dB (ID rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-GTF-ICA</td>
<td>67.55</td>
<td>36.17</td>
<td>19.15</td>
</tr>
<tr>
<td>MFCC</td>
<td>61.17</td>
<td>32.45</td>
<td>5.85</td>
</tr>
<tr>
<td>LPCC</td>
<td>57.47</td>
<td>25.53</td>
<td>10.11</td>
</tr>
<tr>
<td>PLP</td>
<td>51.60</td>
<td>21.81</td>
<td>6.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>20 dB (ID rate)</th>
<th>10 dB (ID rate)</th>
<th>5 dB (ID rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-GTF-ICA</td>
<td>61.17</td>
<td>35.64</td>
<td>22.87</td>
</tr>
<tr>
<td>MFCC</td>
<td>60.64</td>
<td>12.77</td>
<td>6.38</td>
</tr>
<tr>
<td>LPCC</td>
<td>54.25</td>
<td>16.50</td>
<td>9.57</td>
</tr>
<tr>
<td>PLP</td>
<td>50.00</td>
<td>14.89</td>
<td>9.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>20 dB (ID rate)</th>
<th>10 dB (ID rate)</th>
<th>5 dB (ID rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-GTF-ICA</td>
<td>59.04</td>
<td>31.91</td>
<td>23.95</td>
</tr>
<tr>
<td>MFCC</td>
<td>55.85</td>
<td>13.30</td>
<td>6.91</td>
</tr>
<tr>
<td>LPCC</td>
<td>53.19</td>
<td>13.83</td>
<td>7.98</td>
</tr>
<tr>
<td>PLP</td>
<td>50.53</td>
<td>17.02</td>
<td>6.38</td>
</tr>
</tbody>
</table>

7 Conclusions

This paper introduces a robust voice biometric feature extraction technique for speaker identification systems. It is based on single channel speech input, Gammatone auditory filterbank modelling, and Independent Component Analysis in the frequency domain. The segmented speech signal passes through a Gammatone auditory filterbank to
represent the frequency analysis performance of the auditory periphery. Then the output of each filter-bank is rectified and compressed to simulate the behaviour of inner hair cells. Then, the resulting signal is transformed to the frequency domain. Finally the feature matrices are developed by using ICA. The resulting feature efficiently represents the statistical structure of the speech signal and captures the correlation between different Gammatone frequency bands, which denote the distribution of individual speakers.

Speaker identification experiments have been carried out to test the proposed feature in clean and noisy-mismatched environments. The results show that the proposed feature is more robust to different kinds of additive noises, especially for low SNR level, compared to the traditional features, such as LPCC, MFCC and PLP.

The proposed F-GTF-ICA technique has been also benchmarked against the commonly used features using speech corpus recorded in real environments, where speech is tested live. Two environments were taken into the consideration; a non-reverberant room and a normal office room. Our results show that in both environments, the proposed feature can achieve the best identification rate for clean or noisy-mismatched conditions. This shows that the F-GTF-ICA feature matrix not only contains the rich speaker attribute information but also minimises the effect of the additive noise and channel mismatch. At the same time, we notice that the identification rates drop in the normal office room compared with those in the non-reverberant whisper room, due to the reverberation effect.

References


Voice biometric feature using Gammatone filterbank and ICA


