Gammatone Auditory Filterbank and Independent Component Analysis for Speaker Identification

Yushi Zhang *Waleed H. Abdulla*

Department of Electrical and Computer Engineering
The University of Auckland, Private Bag 92019, New Zealand
yzha104@ec.auckand.ac.nz w.abdulla@auckand.ac.nz

ABSTRACT

Feature extraction is the key procedure when aiming at robust speaker identification. The most commonly used feature extraction techniques work successfully only in clean or matched environments. Accurate speaker identification is made difficult due to a number of factors, with handset/channel mismatch and environmental noise being the most prominent. This paper presents a novel technique which based on Gammatone filterbank (GTF) and independent component analysis (ICA). The presented method first relies on the Gammatone filterbank to emulate the human cochlea frequency resolution. By using ICA, it extracts the dominant components from these frequency banks. The extracted features emphasis the difference in the statistical structures among the speakers, which can model the distribution of the individuals. Compared to the commonly used techniques, such as linear predictive cepstral coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC) and perceptual linear predictive (PLP), the proposed method is more robust to additive noises and yields higher recognition rate in mismatch environments in a text-independent speaker identification system.

Index Terms: Speaker Identification, Speaker Recognition, Gammatone filterbank, Independent Component Analysis.

1. INTRODUCTION

A main focus in state-of-art automatic speaker identification systems is finding efficient features for speech signal. So far, the short-time spectral analysis has taken the leading role, such as linear predictive cepstral coefficients (LPCC) [1], Mel-frequency cepstrum coefficients (MFCC) [2] and perceptual linear predictive (PLP) [3] have been shown fairly good performance while used in conjunction with Gaussian Mixture Models (GMM) [4]. However, a major deficiency in speaker identification system is the lack of robustness in mismatched training and testing environments. The mismatch in acoustic characteristics between speech signals produced by training speakers and those testing speakers has been causing serious performance degradation for speaker identification system. Meanwhile, modern speech enabled applications require operation on signal of interest contaminated by high level of noise. The spectral based features are sensitive to various corrupted acoustic conditions and easily distorted by additive noises. Therefore, these situations urge the demand for a greater robustness in estimation of the speech parameters for mismatch environments and low environmental signal-to-noise ratio (SNR) level. In our approach, two assumptions have been taken into consideration to solve the problem. The first assumption is based on that human auditory discrimination typically manifests itself by capability of audio separation in frequency. In the human 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 [6]. Then in Gammatone filterbank modelling, bandpass filters are designed to resemble the characteristics of the frequency selectivity of the basilar membrane. The second assumption is related to the statistical separation power of the human auditory system. Recently, independent component analysis has been shown highly effective in extracting features from a set of observed speech signals by reflecting the statistical structure of the observed signals [7-9]. ICA assumes that the speech signal can be decomposed into basis functions and coefficients. The basis functions of speech maximize the amount of information in the transformed domain, so that the adapted individual basis functions obtained by ICA can be used as features for speaker identification. However, the relevance of ICA features is not entirely transparent and the relation to the auditory system features, that are specific to a speaker, is not clear [5].

In this paper, we apply ICA to speech signals after they pass through a Gammatone filterbank in order to analyze its intrinsic characteristics within the different human perceptual frequency bands and to obtain a new set of features for automatic speaker identification. The extracted features not only represent the statistical structure within Gammatone frequency bands but also capture correlations among these frequency bands specific to the given speaker. Since ICA leads a highly efficient representation of the observed speech signal, the ICA features ignore the effect of the mismatch environments and additive noises. We compared the GTF-ICA features with LPCC, MFCC and PLP by the text-independent speaker identification system on the TIMIT speech corpus and NOISEX-92 noise database. The results prove that the proposed features are more robust to mismatch environments and additive noise and achieve the better identification rate.

2. GTF-ICA FEATURES EXTRACTION

Signal processing front end for extracting the feature set is an important stage in any speaker identification system. In this section, we propose one more successful technique to extract the feature set from a speech signal for speaker identification systems. This feature is based on Gammatone filterbank and

independent component analysis. Idea of the proposed front end is illustrated in Fig. 1.

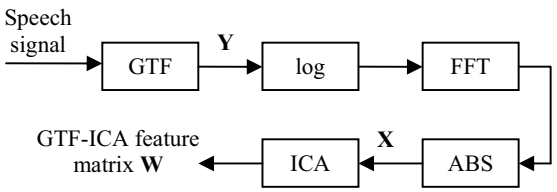

Figure 1: Block diagram of GTF-ICA feature extraction

2.1 Gammatone auditory filterbank processing

In our approach, we first pass speech signals into the Gammatone filterbank. GTF modelling is a physiologically based strategy followed in mimicking the structure of the peripheral auditory processing stage [10]. 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 high resolution of filters can be approximated by specifying certain overlapping between the contiguous filters. The impulse response of each filter follows the Gammatone function shape. And the bandwidth of each filter is determined according to the auditory critical band (CB). The CB is the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea path [10]. The frequency impulse response of a 16-channel filterbank, covering 100-8000Hz band, is shown is Fig. 2.

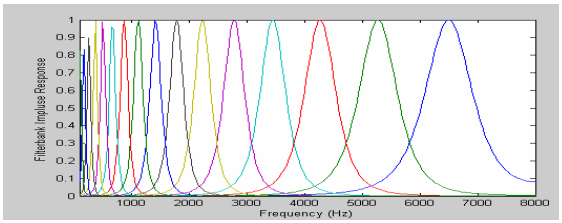

Figure 2: Frequency response of a 16-channel Gammatone filterbank

From Fig. 2., 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 [10].
Assume the number of 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.

2.2 Learning ICA speaker basis functions

The aim of the logarithm is to reduce the dynamic range of filterbank outputs so that ICA method can easily capture the statistic structure of the signal. The Fourier transform of the resulting filterbank outputs are taken from the fact that human aural discrimination typically manifests itself by capability of audio separation in frequency domain. After that, absolute values are taken from the frequency spectra, since the phase information has no effect on the speaker identification system similar to human ear which is phase insensitive [11]. To extract independent feature vectors from speech signal, ICA algorithm is applied to the observation **X**. ICA assumes that the observation **X** is a linear mixture of the independent components $\mathbf{s_i}$.

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{N}} \mathbf{a_i s_i} \qquad (1)$$

where **A** is a $N \times N$ scalar square matrix, denotes the mixing matrix, and the column vector $\mathbf{a_i}$'s are called the basis functions generating the observed signal, whereas $\mathbf{W}=\mathbf{A}^{-1}$ refers to ICA filters that transform the signals into independent activations or source components.

$$\mathbf{S} = \mathbf{W} \cdot \mathbf{X} \qquad (2)$$

The objective of ICA is to infer both the unknown sources $\mathbf{s_i}$ and the unknown basis functions **A** or **W** from the observation **X**. We use the maximizing negentropy learning rule to update the basis function **W**. The details of the derivation can be seen in [12]:

$$\mathbf{w} \leftarrow \mathbf{E}\left[\mathbf{z}\mathbf{g}(\mathbf{w}^{\mathbf{T}}\mathbf{z})\right] - \mathbf{E}\left[\mathbf{g}'(\mathbf{w}^{\mathbf{T}}\mathbf{z})\right]\mathbf{w} \qquad (3)$$

$$\mathbf{w} \leftarrow \mathbf{w} / \left\|\mathbf{w}\right\| \qquad (4)$$

where **z** is the whiten data of **X** and **g** is a function defined by $\mathbf{g}(y) = \tanh(ay)$, $1 \le a \le 2$.
Since the extracted basis functions $\mathbf{w_i}$'s from **W** are extracted in frequency domain, they capture the correlations between frequencies. These correlations can be considered as functions of speaker's glottal or nasal shape [1]. Therefore the GTF-ICA feature matrix **W** is specific to individuals. Meanwhile, ICA leads a highly efficient representation of the speech signal. It does not only decocrrelate the second order statistics but also reduce the higher-order statistical dependencies. Hence it captures the main and essential variabilities of the speech signal, such as gender, accent, age, speech rate and phones realizations. On the other hand ICA ignores the other speech variabilities, such as environments mismatch and additive noise. As a result, GTF-ICA feature matrix reflects the given speaker's attributes, and at the same time, reduces the impact of the mismatch environments and noise on the speech signal. Also the feature matrix represents the statistical structure of the speech signal in different frequency bands. And these frequency bands are taken from Gammatone filterbank which is designed to imitate the frequency resolution of human hearing. Hence this new feature technique also introduces concepts of human aural system to the processing, which humanizes the speaker identification system and makes the system more reliable.

3. SUPERVISED IDENTIFICATION OF SPEAKERS

From the above processing, we can acquire a GTF-ICA feature matrix specific to a given speaker. And this feature matrix denotes the distribution of the speaker. To apply this feature matrix to speaker identification system, we utilize a new pattern classification method to identify the speaker rather than the commonly used classification techniques, such as Euclidean distance measure, dynamic time warping (DTW) [1] and maximum likelihood estimation [4].

The idea of the new algorithm is that the basis functions **W** are estimated from the observation **X** so that the random variables $\mathbf{s_i}$'s are as independent as possible. Assume that feature matrix $\mathbf{W}^{train}$ are trained from a given speaker's speech data. Test data from a particular speaker will induce a similarly low degree of independence when test data is projected on this trained feature matrix. However, if test data from a different speaker is used, her/his data is unlikely to produce a similarly low degree of independence [5]. The feature matrix is not designed to minimize independence on data coming from a speaker characterized by a different correlation structure in the frequency domain. Therefore, the identification score Γ is defined as:

$$\mathbf{S}^{test} = \mathbf{W}^{train} \cdot \mathbf{X}^{test} \quad (5)$$

$$\Gamma_{\mathbf{W}^{train}}(\mathbf{S}^{test}) = \sum_{i<j}^{N} \left|\mathbf{r_{ij}}\right|^{\beta} \quad (6)$$

where $\mathbf{r_{ij}}$ is the normalised mutual information between random variables $\mathbf{s_i}$ representing independent components, and $\boldsymbol{\beta}$ is a positive constant. In our approach, for the sake of simplification, we measure the cross-correlation of $\mathbf{s_i}$ instead of calculating the mutual information between them, since the low degree of independence induces low degree of correlation. And $\boldsymbol{\beta}$ is chosen empirically equate to 2.

For speaker identification, a group of **M** speakers **M**=[1,2,…M] is represented by their GTF-ICA feature matrices $\mathbf{W}_1^{train}$, $\mathbf{W}_2^{train}$, … $\mathbf{W}_M^{train}$. The identity of the test speaker is determined by finding the minimum value of Γ :

$$\hat{\mathbf{M}} = \arg\min_{\mathbf{K}}(\Gamma_{\mathbf{W}_K^{train}}(\mathbf{S}^{test})) \qquad 1 \le \mathbf{K} \le M \quad (7)$$

where $\hat{\mathbf{M}}$ is the expected identity of the test speaker, and Γ is given in equation (6).

4. EXPERIMENTAL RESULTS

Text-independent speaker identification tasks were carried out using TIMIT and NOISEX-92 speech databases to evaluate the performance of the proposed algorithm. TIMIT is a noise free speech database recorded using a high quality microphone sampled at 16 KHz. TIMIT contains utterances of 630 speakers from 8 different dialects of spoken English, and for each speaker there are total of 10 sentences arranged in 3 categories (dialect calibration, random contextual variant and phonetically compact sentences). NOISEX-92 is a noise database which provides various noises recorded in real environments. Both of these databases are standard databases commonly used in benchmarking speech processing systems.

In our proposed system, 100 speakers (12 or 16 speakers were randomly selected from each dialect) from TIMIT were used. The period of training sentences is about 18 seconds and of testing sentences is 5 seconds. 30-channel Gammatone filterbank was adopted, since that can best characterise the human aural processing for speech signal sampled at 16 KHz.

To compare the GTF-ICA algorithm to other commonly used feature extraction techniques, we also generated a baseline system based on 24 orders LPCC, MFCC, PLP (without using the delta coefficients) and 32 components GMMs.

The identification rate is defined by:

$$\text{identification rate} = \frac{N_{correct}}{N_{total}} \times 100\% \quad (8)$$

where $N_{correct}$ denotes the number of the correctly identified speakers and N_{total} is the total number of speakers used to be identified. Some cases were considered in our experiments to investigate the robustness of GTF-ICA feature matrix to mismatch and noisy environments.

4.1 Speaker identification in clean environment

In this experiment, we tested our proposed algorithm in an ideal situation: both training and testing sentence were recorded in clean environments. Table 1 lists the identification results of different feature extraction techniques.

Table1: Identification rate (%) for different feature extraction techniques

	GTF-ICA	LPCC	MFCC	PLP
rate	97.0	96.0	97.0	90.0

Table 1 shows that the best result was obtained by the GTF-ICA and MFCC methods. Meanwhile the same speakers were misidentified by using both of these two methods. That proves the GTF-ICA feature matrix is similar to MFCC, and it efficiently represents the variability of speaker and denotes the distribution of individual.

4.2 Feature investigation in noisy environment

In this experiment, we investigated the effect of noise on different feature extraction techniques. The simulation was carried on as following: Reference feature vectors (or matrices) were generated from a speech signal, after that, various noises were added to this speech signal to produce noisy feature vectors (or matrices). Then we compared them with the reference (clean) feature to find the similarity between them. The similarity of two vectors (or matrices) was measured by calculating the cross-correlation coefficient between them. The higher value of cross-correlation coefficient means higher similarity. The additive noises are white Gaussian noise and some other colour noises from NOISEX-92 database, i.e., factory noise, vehicle interior noise, and babble noise. The simulation results are summarized in figures 3 to 6.

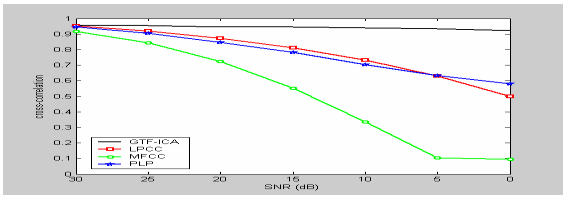

Figure 3: Cross-correlation between clean and noisy features produced in Gaussian white noisy environment

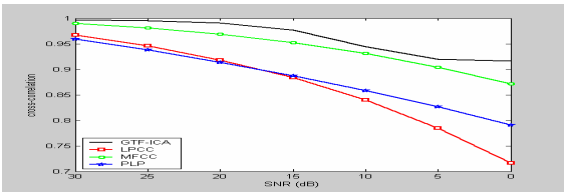

Figure 4: Cross-correlation between clean and noisy features produced in factory noisy environment

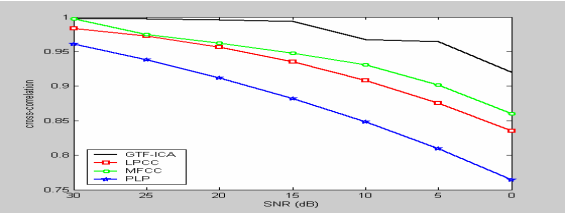

Figure 5: Cross-correlation between clean and noisy features produced in vehicle interior noisy environment

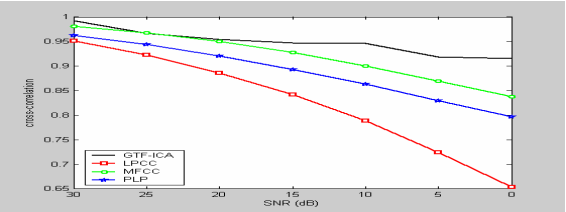

Figure 6: Cross-correlation between clean and noisy features produced in babble noisy environment

It is evident that the GTF-ICA feature matrices produced in noisy environments are similar with those produced in clean environment. However, LPCC, MFCC, and PLP features vary largely when the signal contaminated by noises. Therefore, the GTF-ICA feature matrix is less sensitive to additive noises

4.3 Feature investigation in mismatch environments

In this experiment, the effect of mismatch environments was investigated. Two additive noises, white and factory noises with various SNR levels were added to the training and testing speech utterances, respectively. The main reason we chose these two types of noise is that the conventional techniques work extremely poor in this situation.The simulation results are shown in Table 2.

Table 2: Identification rate (%) in mismatch environments

rate	tr=clean te=20dB	tr=clean te=15dB	tr=25dB te=20dB	tr=20dB te=20dB
GTF-ICA	75.0	66.0	24.0	14.0
LPCC	70.0	48.0	16.0	9.0
MFCC	71.0	56.0	19.0	12.0
PLP	42.0	17.0	4.0	4.0

where 'tr' denotes the training sentences with additive white Gaussian noise and 'te' denotes the testing sentences with additive factory noise respectively. Apparently, the best performance is achieved by using GTF-ICA method, which proves that our proposed algorithm is more robust to mismatch environments compared with all the other commonly used feature extraction techniques.

5. CONCLUSIONS

In our work, we proposed a new feature extraction method, which based on Gammatone filterbank modelling and independent component analysis. The extracted feature efficiently represents the statistical structure of the speech signal, and captures the correlation between different Gammatone frequency bands. The proposed feature does not only denote the distribution of individual speakers but also minimize the effect of the additive noise and mismatch environments. In comparison to the conventional LPCC, MFCC and PLP techniques, our new algorithm is more robust to additive noises and achieves better identification performance in mismatch environments. However, the computational cost of the new algorithm is higher than the conventional techniques due to the need to compute the cross-correlation between the components. In our future research, we will be focusing on finding ways to reduce this cost.

6. ACKNOWLEDGEMENT

This research is partially funded by UARC grant 3603819.

7. REFERENCES

1. Rabiner, L. R. and Juang, B. H., *Fundamentals of speech recognition*. Prentice-Hall signal processing series, Englewood Cliffs, N.J. PTR Prentice Hall 1993.
2. Gish, H. and Schmidt, M., *Text-independent speaker identification.* Signal Processing Magazine, IEEE, **11**(4): p. 18-32, 1994.
3. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* Journal of the Acoustical Society of America, **87**(4): p. 1738-52, 1990.
4. Reynolds, D. A. and Rose, R. C., *Robust text-independent speaker identification using Gaussian mixture speaker models.* Speech and Audio Processing, IEEE Transactions on, **3**(1): p. 72-83, 1995.
5. Rosca, J. and Kofmehl, A., *Cepstrum-like ICA Representations for Text Independent Speaker Recognition.* 4th Int. Independent Component Analysis and Blind Signal Separation, Nara, Japan, p. 1-4, 2003.
6. Lee, J. H., Jung, H. Y., Lee, T. W. and Lee, S. Y., *Speech feature extraction using independent component analysis. Acoustics, Speech, and Signal Processing, ICASSP '00 Proceedings, IEEE International Conference,* 2000.
7. Jang, G. J., Lee, T. W. and Oh, Y. H., *Learning statistically efficient features for speaker recognition. Acoustics, Speech, and Signal Processing, (ICASSP '01), IEEE International Conference,* 2001.
8. Lee, T. W. and Jang, G.-J., *The statistical structures of male and female speech signals. Acoustics, Speech, and Signal Processing, (ICASSP '01), IEEE International Conference on*. 2001.
9. Lee, H. J., Lee, T. W., Jung, H. Y. and Lee, S. Y., *On the Efficient Speech Extraction Based on Independent Component Analysis.* Kluwer Academic Publisher, **15**(3): p. 235-245, 2002.
10. Abdulla, W. H., *Auditory Based Feature Vectors for Speech Recognition System*, *Advance in Communication and Software Technologies*, N. E. Mastorakis & V.V. Kluev, Editor. WSEAS Press. p. 231-236, 2002,
11. Klevans, R. L. and Rodman, R. D., *Voice recognition*. The Artech House telecommunications library. Boston, Artech House, 1997.
12. Hyvärinen, A., Karhunen, J. and E. Oja, *Independent component analysis*. New York, J. Wiley, 2001.