interspeech 2008 incorporating SST 08

Robust Speaker Identification Using Cross-correlation GTF-ICA Feature

Yushi Zhang, W. H. Abdulla

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

10.21437/Interspeech.2008-506

Abstract

Robust feature for speaker identification in noisy environments is proposed. This method is inspired by the human binaural auditory system. A pair of microphones is used to replicate human ears in the processing. Cross-correlation processing is taken of the microphone outputs after Gammatone bandpass filtering, rectification and compression. ICA is then applied to the real cepstrum of the correlated waveform to extract the dominant components from each frequency band. The resulting feature emphases the difference in the statistical structures among speakers. Compared to the commonly used MFCC techniques, the proposed method is more robust to background noises and provides higher identification rate in real noisy environments for text-independent speaker identification systems. A specially prepared noisy speech corpus was used to gauge the performance of the proposed feature.

Index Terms: speaker identification, Gammatone Filterbank, Independent Component Analysis, binaural model

1. Introduction

The major deficiency in speaker identification systems is the lack of robustness in noisy environments. The commonly used spectral based features, such as Mel-frequency cepstrum coefficients (MFCC) [1], are sensitive to various corrupted acoustic conditions and easily distorted by background noises. In our proposed approach, two factors have been taken into consideration. The first factor is based on the human binaural system, which can be modeled by a pair of input microphones (our ears). The system includes Gammatone auditory bandpass filtering of speech signal, half-wave rectification, and A-law compression to model the effects of the auditory system periphery. Then the two-channel resulting speech waveforms are combined by cross-correlation processing to minimize the speech-unrelated noisy effect. The second factor is based on the signal separation power of the human auditory system. Independent component analysis (ICA) has been shown highly effective in separating signals and extracting features from a set of observed speech signals by reflecting the statistical structure of the observed signals [3,4]. The proposed feature performance has been compared with MFCC features using a text-independent speaker identification system trained on clean TIMIT speech corpus and tested by noisy speech corpus specially prepared to replicate real environments, as explained in section 5. The results prove that the proposed feature is more robust to noises and achieves better identification rate.

The paper is structured as follows; the proposed method is described in section 2. Then the effectiveness and robustness against background noises of the proposed feature are investigated in section 3. Section 4 introduces the dominant cross diagonal feature extraction. In section 5, a series of experiments on the task of speaker identification under various noisy environments is conducted to evaluate the performance of the proposed algorithm. Finally, conclusions are given in section 6.

2. The proposed feature matrix

In this section, we detail the components of the proposed technique. It is based on the human binaural system as illustrated in Fig. 1 and explained in the following sections.

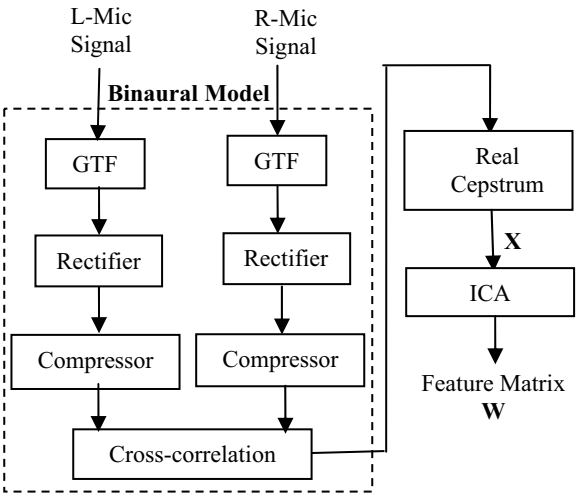

Figure 1: *Block diagram of GTF-ICA feature extraction*

2.1. Binaural models

This model carefully imitates the functions of human hearing according to physiological and psychoacoustic clues.

A pair of microphones is used to replicate human ears in the system. Then the acquired signals from these two microphones are passed to the Gammatone filterbanks (GTF). GTF models the cochlea by a bank of overlapping bandpass filters. The impulse response of each filter follows the Gammatone function shape [2]. Each bandpass filter of the GTF is followed by a half-wave rectifier and an A-law compressor to simulate the behaviour of inner hair cells. The half-wave rectifier accurately simulates transduction from the movements of basilar membrane to neural activity in the inner hair cells. The hairs at the tops of these inner cells are displaced when the basilar membrane moves up and down. As a result, the auditory nerve encodes a half-wave rectified version of the stimulus, because action potentials are only initiated by the movement of hairs in one direction. The aforementioned measurements also show a compressive response. Therefore, we apply A-power-law compressor to

the rectified signals to reduce the dynamic range of the signal.Assume that the speech signal and background noises received at the two microphones are uncorrelated. Then the cross-correlation of these two subband outputs boosts the speech signal component and reduces the effect of noisy component. The cross-correlation of two signals can be considered as another representation of the signal in time domain, which may also carry the spectral energy information from each frequency band.

2.2. Cepstral analysis

The cepstral analysis of the cross-correlation output is performed due to the fact that human aural discrimination typically manifests itself by capability of audio separation in the cepstral domain. In addition, cepstral coefficients are more robust than spectral coefficients in noisy environments. The real cepstrum is adopted since the phase information has no effect on the speaker identification system similar to the human ear, which is phase insensitive [1].

2.3. Learning ICA speaker basis functions

ICA method is then applied to the cepstrum observation **X** (in Figure 1) to extract independent feature vectors. ICA assumes the observation **X** as a linear mixture of independent components $\mathbf{s_i}$.

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

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

where **N** is number of components and **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. 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 is depicted in [5].

The extracted basis functions $\mathbf{w_i}$ capture the certain correlations among the frequencies present in the spectral based representation of a speech signal. This is achieved by ICA in the form of linear combinations of basic filter functions specific to each person. These correlations can be considered as functions of a 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 decorrelate the second order statistics but also reduces the higher-order statistical dependencies. It ignores the other speech variabilities, such as additive noise. The two-input cross-correlation processing and cepstral analysis improve the robustness against uncorrelated noises. As a result, the feature matrix represents speaker's attributes, and at the same time, reduces the impact of noise on the speech signal.

3. Feature matrix investigation

In this section, the effectiveness and robustness of our proposed feature matrix are investigated.

3.1. Feature evaluation analysis

In this experiment, the proposed features are extracted from speakers' utterances recorded in clean environment. The feature from now on will be called two channel GTF-ICA (TC-GTF-ICA). Two scenarios are considered: the same speaker speaks different texts (case1), and different speakers speak the same text (case2). Then the similarity between the features is measured by calculating the cross-correlation coefficients. The cross correlation coefficient is a standard method of estimating the degree to which two series are correlated. The value of the cross-correlation coefficient is between 0 and 1, and a higher value means a higher similarity. Table 1 summaries the resulting cross-correlation coefficient and the correlation ratio between the cross-correlation coefficients produced by same speaker with different texts (case1) and that produced by different speakers with same text (case2). For comparison, the performance of the MFCC feature is also included in the table. Here T-G-I denotes TC-GTF-ICA for the sake of simplification.

Table 1. *Cross-correlation coefficients between TC-GTF-ICA and MFCC features extracted from two speakers' utterances*

		2 s	3 s	4 s	5 s
Case1	T-G-I	0.7004	0.8103	0.8328	0.8553
	MFCC	0.6478	0.8023	0.8437	0.8748
Case2	T-G-I	0.4204	0.4856	0.5193	0.5394
	MFCC	0.7583	0.7732	0.8274	0.8489
Corr ratio	T-G-I	**1.6660**	**1.6687**	**1.6037**	**1.5857**
	MFCC	0.8543	1.0376	1.0197	1.0305

Apparently even when the same speaker speaks different texts, the TC-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 TC-GTF-ICA feature matrix contains more speaker related information than linguistic information, which is desired in a speaker identification system. 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 TC-GTF-ICA. However, MFCC also produces high degree of similarity when different speakers 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. Comparing the correlation ratio obtained by TC-GTF-ICA and MFCC, the former has higher value and must be more suitable to speaker identification.

3.2. Robustness of feature investigation

In this experiment the robustness of the TC-GTF-ICA feature matrix against background noises is investigated. A reference feature matrix is calculated from a 5-second speech utterance recorded in a clean environment. Then different background noises (factory and destroyer noises) with various SNR levels were produced by loudspeakers while recording this speech utterance. The noisy feature matrices are calculated using the proposed technique. Then we compare them with the reference (clean) feature to investigate how much distortion caused by background noises. The feature distortion is defined as the equation 3, where, $\rho_{(clean,noisy)} \in \{0,1\}$ is the cross-correlation coefficient between clean and noisy features. Therefore feature distortion is also between 0 and 1. A large value denotes a large distortion. The following

figures plot the resulted distortion caused by the two background noises.

$$feature \quad distortion = 1 - \rho_{(clean,noisy)} \tag{3}$$

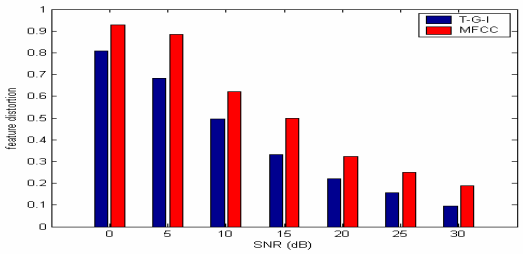

Figure 2: *Feature distortion caused by mixing speech with factory noise*

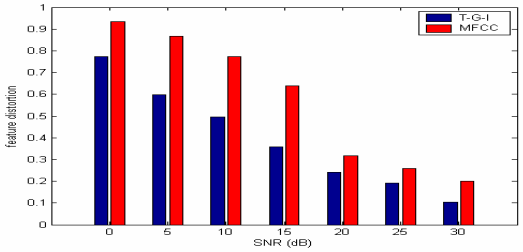

Figure 3: *Feature distortion caused by mixing speech with destroyer operations noise*

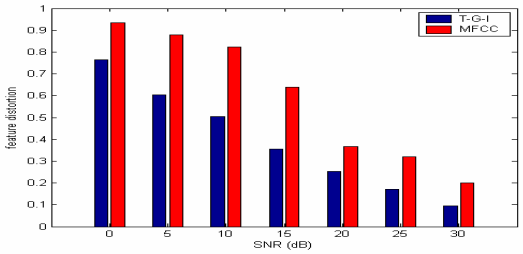

Figure 4: *Feature distortion caused by mixing speech with factory and destroyer operations noises*

It is evident that the TC-GTF-ICA feature has less distortion compared with MFCC features when the speech signal is contaminated by background noises. That proves the proposed feature is more robust to noises.

4. Cross diagonal feature extraction

From section 2, we conclude that we can acquire a TC-GTF-ICA feature matrix specific to a given speaker. Fig. 5 plots the 3-D distribution of a TC-GTF-ICA feature matrix.

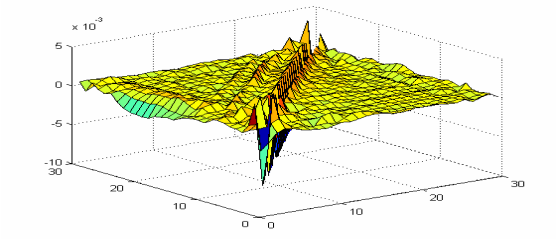

Figure 5: 3-D surface of a TC-GTF-ICA feature distribution

As can be seen, the most noticeable elements are located at the diagonal area of the matrix. This led us to select the cross diagonal elements only as speech feature for speaker identification to reduce the computation load.

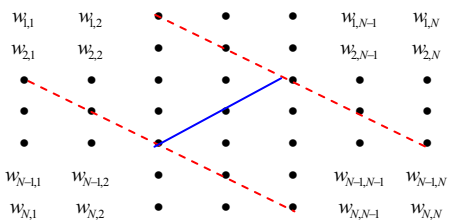

Figure 6: Diagram of a TC-GTF-ICA feature matrix

Fig.6. shows a TC-GTF-ICA feature matrix and the boundaries of the effective elements. We add the diagonal elements between the two dash lines (red line) positioned on the solid line (blue line) direction, and take the sum as a speech feature vector used for further processing. We have found through using the F-ratio figure-of-merit that summing 5 diagonal elements gives the best performance. F-ratio is a measure that can evaluate the effectiveness of the feature coefficients and has been widely used as a figure-of-merit for feature selection in speaker recognition applications [2]. In the experiment, 188 TC-GTF-ICA feature matrices produced from 188 different speakers from TIMIT database were used. The F-ratio decreases when more than 5 elements included. That is because the additional elements may contain some speaker irrelevant information, which may be language information, or noisy information. Therefore, in our proposed method, the sum of five cross diagonal elements is used as feature for speaker identification. Our experimental results also verify that the best identification result can be achieved while using only five cross diagonal elements.

5. Experimental results

Text-independent speaker identification tasks are carried out in real environments to evaluate the performance of the proposed feature. In our experiment set up, a source loudspeaker was producing the TIMIT database while two other loudspeakers were generating “factory” or “destroyer” noises from NOISEX-92 database. 188 speakers (88 females and 100 males were randomly selected from 8 dialect) from TIMIT are used. We test our proposed algorithm in a real noisy-mismatched situation: for each speaker, the source loudspeaker produced 15 seconds training utterance. Two microphones emulating human ear stood in front of the source loudspeaker to record speech signals. All the training utterances were recorded in clean environment without any background noises or room reverberation. A photo of the room is shown in Fig.7. While testing utterances were recoded in the same non-reverberant room. The source loudspeaker produced 5 seconds testing utterance while noise loudspeakers were producing the background noises at different levels. Four cases are taken into consideration.

Case 1: The source loudspeaker is active while the noise loud speakers are off. Therefore both training and testing utterances are recorded in a noise free environment.

Case 2: The source loudspeaker and noise loudspeaker producing factory noise are active. The measured signal-to-noise ratio (SNR) is set to be 20dB, 10dB, and 5dB by adjusting the noise level.

Case 3: The source loudspeaker and noise loudspeaker producing destroyer noise are active. The measured SNR is around 20dB, 10dB, and 5dB.

Case 4: All loudspeakers of source and noises are active. The measured signal to noise ratio was again adjusted to be about 20dB, 10dB, and 5dB.

Figure 7: speech recording in non-reverberation room

The recorded speech signal is first segmented into 30ms frames with 50% overlap, and a GTF-ICA feature matrix is extracted using our proposed method from each frame. 30-channel Gammatone filterbank is adopted, which best characterises the human aural processing for speech signal sampled at 16 KHz. After that, five cross diagonal elements of the feature matrix are added to form the feature vectors. In our speaker identification system, these vectors are used in conjunction with 32 components Gaussian Mixture Models (GMM)[1]. For comparison, we generated a baseline system using MFCC with order 24 and 32 components GMMs. We also apply Cepstral Mean Subtraction (CMS) compensation to the proposed features to improve the robustness against the mismatched acoustic conditions. The reason for CMS compensating the channel mismatch has been well known [1]. During our speech signal recording, the recording room can be considered as a recording channel and the impulse response of the recording room varies due to the geometry, furniture and decoration of the room. The identification rate results based on the four cases are shown in Table 2.

Table 2. *Identification rates (%) in a non-reverberation room*

rate		T-G-I	T-G-I (CMS)	MFCC	MFCC (CMS)
Case1		96.81	**96.81**	96.81	96.81
Case2	20dB	76.60	**79.79**	65.43	75.53
	10dB	48.49	**54.26**	36.70	47.87
	5dB	31.91	**37.17**	11.17	21.81
Case3	20dB	75.53	**77.13**	65.96	72.34
	10dB	50.53	**56.91**	21.81	24.47
	5dB	39.89	**48.40**	12.77	18.61
Case4	20dB	73.40	**75.53**	61.17	70.21
	10dB	49.47	**52.13**	17.02	24.47
	5dB	36.70	**47.87**	11.70	22.34

Table 2 shows that the same identification rate is obtained by the TC-GTF-ICA and MFCC methods in the clean environment. It has been noticed that the same speakers were misidentified by both two methods. This proves that the TC-GTF-ICA feature matrix efficiently represents the variability of speakers and denotes the distribution of individual. CMS compensation has no effect on the system performance for the clean environment.

In noisy-mismatched environment, TC-GTF-ICA outperforms the baseline system, even though MFCC is compensated with CMS algorithm. With Cepstral mean subtraction, identification performances of TC-GTF-ICA and MFCC are improved, since CMS is an efficient method of removing the distortion caused by acoustic conditions and channel mismatch. With CMS compensation, TC-GTF-ICA algorithm has the best identification performance in noisy-mismatched environments. This proves that our algorithm is more robust to additive noise compared to the MFCC feature.

The speech corpus recording for testing was also repeated in a normal office room to investigate the room reverberation effect on the proposed method. Table 3 summaries the resulting identification rate. It is evident that the identification rates of TC-GTF-ICA and MFCC decrease when the room reverberation is taken into consideration. Comparing the identification performances degradation of the two algorithms, the TC-GTF-ICA rate decreases 3.90% on average, while MFCC has 5.03%. That proves TC-GTF-ICA feature is more robust against reverberation than MFCC.

Table 3. *Identification rates (%) in a normal office room*

rate		T-G-I	T-G-I (CMS)	MFCC	MFCC (CMS)
Case2	20dB	72.34	**75.53**	61.17	70.74
	10dB	44.68	**48.94**	32.45	42.02
	5dB	25.53	**30.32**	5.85	12.77
Case3	20dB	69.68	**75.00**	60.64	68.62
	10dB	45.74	**51.60**	14.89	20.21
	5dB	35.11	**40.43**	7.44	12.23
Case4	20dB	68.62	**71.28**	55.85	65.43
	10dB	43.62	**47.87**	13.30	20.75
	5dB	34.57	**41.49**	6.91	15.43

6. Conclusions

We have proposed a robust feature for speaker identification inspired by human binaural system. The extracted feature efficiently represents the statistical structure of the speech signal, and captures the correlation between the Gammatone frequency bands of two channels. By using two-microphone cross-correlation processing, cepstral analysis and independent component analysis, the proposed feature does not only denote the distribution of individual speakers but also minimize the effect of the background noise. In comparison to the conventional MFCC techniques, our algorithm is more robust to background noises and room reverberation, and achieves better identification performance in real noisy-mismatched environments. Effectiveness and robustness against background noise of the proposed feature have been evaluated using a speech corpus prepared specifically to emulate real noisy environments.

7. Acknowledgement

This project is supported by UARC grant 3607215.

8. 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. 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.
3. 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.
4. 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.
5. Hyvärinen, A., Karhunen, J. and E. Oja, *Independent component analysis*. New York, J. Wiley, 2001.