Gammachirp Filter Banks Applied in Roust Speaker Recognition Based on GMM-UBM Classifier

Lei Deng and Yong Gao
College of Electronics and Information Engineering, Sichuan University, China

Abstract: *In this paper, authors propose an auditory feature extraction algorithm in order to improve the performance of the speaker recognition system in noisy environments. In this auditory feature extraction algorithm, the Gammachirp filter bank is adapted to simulate the auditory model of human cochlea. In addition, the following three techniques are applied: cube-root compression method, Relative Spectral Filtering Technique (RASTA), and Cepstral Mean and Variance Normalization algorithm (CMVN).Subsequently, based on the theory of Gaussian Mixes Model-Universal Background Model (GMM-UBM), the simulated experiment was conducted. The experimental results implied that speaker recognition systems with the new auditory feature has better robustness and recognition performance compared to Mel-Frequency Cepstral Coefficients (MFCC), Relative Spectral-Perceptual Linear Predictive (RASTA-PLP),Cochlear Filter Cepstral Coefficients (CFCC) and gammatone Frequency Cepstral Coefficeints (GFCC).*

Keywords: *Feature extraction, gammachirp filter bank, RASTA, CMVN, GMM-UBM.*

Received May 9, 2017; accepted June 19, 2019
https://doi.org/10.34028/iajit/17/2/4

1. Introduction

Speaker recognition as a typical application of biometrics is being gradually applied to various fields. Nevertheless, the conditions in real world are not ideal and it always differs from those in laboratory, which could easily result in mismatches between training and testing environments. As a result, the recognition performance will be degraded dramatically. In the existing literature, there have been four main methods to improve the robust speaker recognition [2, 12]: anti-noise feature extraction (extracting the characteristics of speech that are insensitive to the noise), speech enhancement (recovering/estimating the clean signal from its contaminated version), model compensation(m-odifying the parameters of the pure speech model according to the characteristics of ambient noise to compensate for mismatches between training and testing environment) and investigation on human auditory (the auditory characteristics of the human auditory system have stronger noise robustness). In this paper, authors will mainly implement the extraction method based on the anti-noise feature.

The human auditory system has stronger noise robustness and better recognition property under low Signal-to-Noise (SNR) conditions. The study of the auditory system focused on three aspects: experimental studies the auditory system, the auditory system modeling, and applications of the auditory system modeling [11]. The cochlea is the vital organ of the human auditory system and the basement membrane is an important structure of the cochlea. The basement membrane is generally taken as a set of band pass filter bank. The basement membrane as the filter has three characteristics [14]: non-uniform filter bandwidths; asymmetric frequency response of individual filters; level-dependent frequency response of individual filters. The prior studies have successfully engaged many fine features in the human auditory model. Now, the two most common robust features in the speaker recognition system are Mel-Frequency Cepstral Coefficient (MFCC) [16], Relative Spectral-Perceptual Linear Predictive (RASTA-PLP) [8], Cochlear Filter Cepstral Coefficients (CFCC) [14] and Gammatoin Filter Cepstral Coefficients (GFCC) [20]. In particular, MFCC partially considers the auditory characteristics of the human auditory system; however under clean speech conditions, mismatches between the training and testing environments would cause the recognition rate to significantly drop from 100% to 15.6% [4]. In contrast, under clean speech conditions, the recognition rate for MFCC could reach 96%, but when the SNR of the input signal is 6dB, the recognition rate drops to 41.2% [13]. In Tazi *et al.* [20], Further improved the recognition rate by applying the Gammatone auditory filter to extract the speech feature parameters. This design method achieves good results. However, the problem is that the amplitude-frequency response of the Gammatone auditory filter is symmetric of the center frequency and there is no level-dependent characteristic in the Gammatone. Therefore, the characteristics of the basement membrane could not be illustrated. To fix this problem, Irino and Patterson [9] proposed the Gammachirp filter that could improve the simulation of the basement membrane characteristics.

Based on Irino’s work, Abdallah and Hajaiej [1], Salhi *et al.* [19] and Bouchamekh *et al.* [3] all successively utilized the Gammachirp filter bank to extract the auditory feature parameters, and applied it to the speaker recognition system.

On the basis of previous findings from human auditory experiments, this paper proposed an auditory feature extraction algorithm using Gammachip filter, and the following three techniques are applied in the feature extraction: cube root compression [6], RASTA [7], and Cepstral Mean and Variance Normalization Technique (CMVN) [17]. What’s more, author also introduced the application of the GMM-UBM technique for the speaker recognition system. The experimental results have demonstrated that the proposed feature could effectively characterize the human auditory system and result in the higher recognition rate compared to other features.

The overall structure of the speaker recognition system is introduced in section 2. The proposed feature extraction method with Gammachirp filter banks is presented in section 3. It consists in speech pre-processing, in cube-root compression, RASTA filter techniques and CMVN method. Experiment setup is elaborated in section 4, including training set preparation, testing set preparation, and three different experimental setups. Experimental results reported in section 5 show the effectiveness of proposed feature extraction algorithm.

2. Speaker Recognition System Architecture

As the subset of pattern recognition, speaker recognition aims to recognize the object based on the prior knowledge of the object [15]. In this section, a formal description is provided to demonstrate the two fundamental components of the speaker recognition system: training and testing. At the training stage, the speaker’s discriminatory information is extracted via processing a set of clean speech signals. After that, the discriminative information would be used to construct the Gaussian Mixture Model (GMM) of the speaker as an input. At the testing stage, the recognition system has to extract the features from noisy speech signals and compare them against the stored models. Subsequently, the recognition results are recorded according to the match score. In summary, Figure 1 illustrates the architecture of the speaker recognition system.

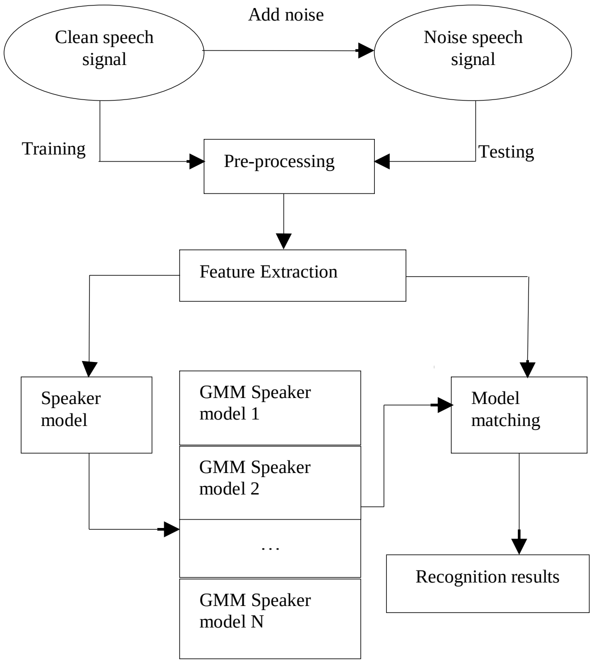

Figure 1. Components of a typical speaker recognition system.

GMM represents a speaker’s feature has become the dominant approach for the speaker recognition system. Reynolds and Rose [18] firstly presented Gaussian Mixture Model-Universal Background Model (GMM-UBM) for the speaker recognition system and GMM-UBM widely used in the speaker recognition systems over the last decade. The basic idea in the GMM-UBM is to derive the hypothesized speaker specific model by adapting the parameters of UBM using the speaker’s utterances and a form of Bayesian adaptation [5]. A formal description is provided to demonstrate the three fundamental components of the GMM-UBM based speaker recognition system: Universal Background Model (UBM) training, Bayesian adaptation of the UBM and speaker recognition. Firstly, two speaker model groups were built for male speaker and female speaker as Figure 2. The UBM is consisting of male UBM and female UBM. Then, the GMM speaker models are derived by adapting the parameters of the UBM using the speaker’s utterances. Also, the testing utterances are used to obtain significant mixture from the UBM. Lastly, the max score is computed by log-likelihood [21].

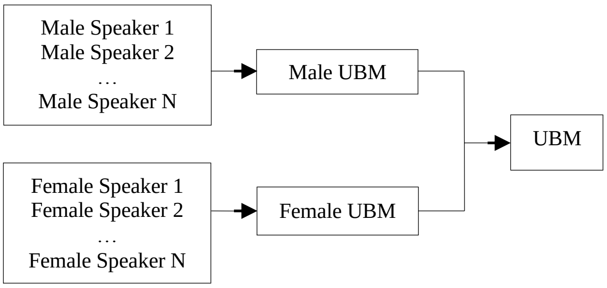

Figure 2. UBM training.

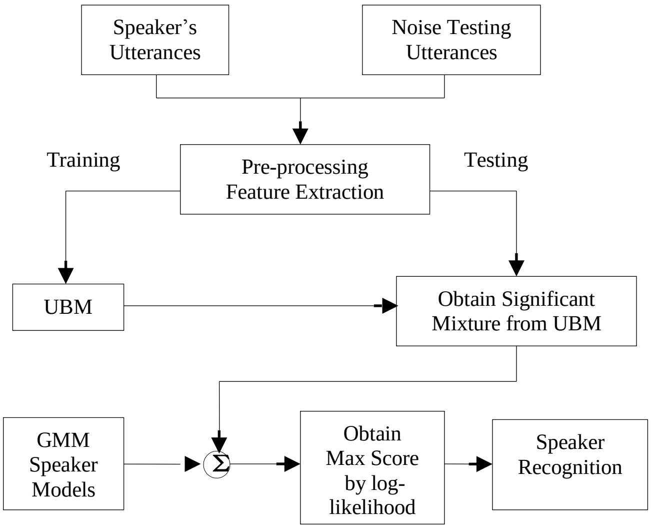
Speaker's Utterances
Noise Testing Utterances
Training
Pre-processing Feature Extraction
Testing
UBM
Obtain Significant Mixture from UBM
GMM Speaker Models
Σ
Obtain Max Score by log-likelihood
Speaker Recognition

Figure 3. Components of a GMM-UBM based speaker recognition system.

The pre-processing includes Optimally Modified Log-Spectral Amplitude Estimator (OM-LSA) [5] speech enhancement algorithm, pre-emphasis, framing, and windowing. Firstly, OM-LSA algorithm is applied to the input speech signal and $y^{(n)}$ is obtained. Secondly, the pre-emphasis is to enhance high frequencies of the signal and it is implemented by a Finite Impulse Response (FIR) filter,

$$H(z)=1-0.95z^{-1} \tag{1}$$

Since the range of the speech signal is from 10 to 30ms which is short in time, the signal could be divided into several frames, where the i^{th} frame has 256 samples. Then, each individual frame is windowed with hamming windows ($w(n)$) in order to minimize the number of signal discontinuities at the beginning and the end of each frame. $x(n)$ is the result of windowing signals.

$$w(n)=\begin{cases}0.54-0.46\times\cos(\frac{2\pi n}{2N-1}) & 0\le n\le N-1\\ 0 & otherwise\end{cases} \tag{2}$$

$$x(n)=y(n)\times w(n) \tag{3}$$

3. Feature Extraction with the Gammachirp Filter Banks

3.1. Gammachirp Auditory Filter

After collecting the physiological and psychological experimental data, Irino found that the auditory mechanism of the inner ear has a nonlinear characteristic. Based on this founding, he proposed the Gammachirp filter bank at first. The time-domain impulse response of the Gammachirp filter bank could be expressed as:

$$g_c(t)=\lambda t^{n-1}\exp(-2\pi bERB(f_r)t)\cdot\exp(j2\pi\cdot f_r t+jc\ln(t)+j\phi) \tag{4}$$

Where λ is the filter gain, n is the filter order and when n=4, it is a good simulation of the human basement membrane, f_r is the center frequency and ϕ is the phase (usually $\phi=0$). $ERB(f_r)$ is the equivalent rectangular bandwidth of an auditory filter at moderate levels (Hz).

$$ERB(f_r)=24.7+0.108f_r \tag{5}$$

In addition, the center frequency of r^{th} filter f_r can be given by the following equation:

$$f_r=-228.7+(f_H+228.7)\exp(\frac{vr}{9.26})\quad(1\le r\le N) \tag{6}$$

$$v=\frac{9.26}{N}\log(\frac{f_H+228.7}{f_L+228.7}) \tag{7}$$

Where f_H is the high cutoff frequency of the filter, f_L is the low cutoff of the filter. N is the number of filters and v is the percentage-overlapping factor. Note that v is also used to represent the overlapping percentage between adjacent filters.

Moreover, the chirp factor c [10] is a parameter which determined the level-dependent frequency response of individual filters.

$$c=3.38+0.107Ps \tag{8}$$

Where ps is the power of the speech signal. Note that when c=0, the chirp term, $j$cln(t), the Gammachirp function degenerates to the Gammatone function.

The Fourier transform of the Gammachirp could be derived as following:

$$G_C(f)=\int_0^{+\infty}g_c(t)e^{-j2\pi ft}dt=\int_0^{+\infty}\lambda t^{n-1+jc}e^{-2\pi\beta t}e^{j2\pi f_r t+j\phi}e^{-j2\pi ft}dt \tag{9}$$

Where $\beta=b\cdot ERB(f_r)$, $\begin{matrix}u=(2\pi\beta+j2\pi f\text{-}j2\pi f_r)t\\ dt=\frac{1}{2\pi\beta+j2\pi f\text{-}j2\pi f_r}du\end{matrix}$, therefore,

$$G_C(f)=\lambda e^{j\phi}(\frac{1}{2\pi\beta+j2\pi f\text{-}j2\pi f})^{n+jc}\int_0^{+\infty}u^{n-1+jc}e^{-u}du \tag{10}$$

$$G_C(f)=\frac{a}{(2\pi\beta+j2\pi f\text{-}j2\pi f)^{n+jc}} \tag{11}$$

$$\begin{aligned}G_C(f)&=\frac{a}{(2\pi)^{n+jc}(\beta+j(f-f_r))^{n+jc}}\\&=\frac{a}{(2\pi)^{n+jc}(\sqrt{\beta^2+(f-f_r)^2}(\frac{\beta}{\sqrt{\beta^2+(f-f_r)^2}})+j(\frac{(f-f_r)}{\sqrt{\beta^2+(f-f_r)^2}}))^{n+jc}}\end{aligned} \tag{12}$$

Where $(\frac{\beta}{\sqrt{\beta^2+(f-f_r)^2}})^2+(\frac{(f-f_r)}{\sqrt{\beta^2+(f-f_r)^2}})^2=1$,

$\cos\theta(f)=\frac{\beta}{\sqrt{\beta^2+(f-f_r)^2}}$, $\sin\theta(f)=\frac{(f-f_r)}{\sqrt{\beta^2+(f-f_r)^2}}$ therefore,

$$\begin{aligned}G_C(f)&=\frac{a}{(2\pi)^{n+jc}(\beta+j(f-f_r))^{n+jc}}\\&=\frac{a}{(2\pi)^{n+jc}(\sqrt{\beta^2+(f-f_r)^2}(\sin\theta(f)+j\cos\theta(f)))^{n+jc}}\end{aligned} \tag{13}$$

According to Euler's Equation,

$$\begin{aligned} G_C(f) &= \frac{a}{(2\pi)^{n+jc}(\beta + j(f - f_r))^{n+jc}} \\ &= \frac{a}{(2\pi)^{n+jc}(\sqrt{\beta^2 + (f - f_r)^2} e^{j\theta(f)})^{n+jc}} \\ &= a \cdot \left[\frac{1}{\left\{2\pi\sqrt{\beta^2 + (f - f_r)^2}\right\}^n} \cdot e^{-jn\theta(f)} \right] \cdot \left[\frac{1}{\left\{2\pi\sqrt{\beta^2 + (f - f_r)^2}\right\}^{jc}} \cdot e^{c\theta(f)} \right] \end{aligned} \quad (14)$$

$$\theta(f) = arctg\,\frac{f - f_r}{\beta} \quad (15)$$

Where $a = \lambda\Gamma(n + jc)e^{j\phi}, \Gamma(n + jc) = \int_0^{+\infty} t^{n+jc-1}e^{-t}dt$ and a can be treated as constants.

Thus, the amplitude spectrum of the Gammachirp filer can be expressed in terms of the Gammatone as:

$$|G_C(f)| = \frac{1}{\left\{2\pi\sqrt{\beta^2 + (f - f_r)}\right\}^n} \cdot e^{c\theta(f)} = |G_T(f)| \cdot e^{c\theta(f)} \quad (16)$$

Where $|GT^{(f)}|$ is the Fourier magnitude spectrum of the Gammatone filter , $e^{c\Theta(f)}$ is a unit step function. It is noted that when c=0, Equation (16) degenerates to the amplitude spectrum of the Gammatone function $|GT^{(f)}|$.

In Figure 4, it shows the relationship between the three amplitude spectra. $e^{c\Theta(f)}$ is defined as an asymmetric function centered at the asymptotic frequency. Amplitude-frequency response of Gammachirp filter showed a significant asymmetry, which is the most obvious difference between Gammachirp filter and Gammatone filter. Thus, the Gammachirp filter could be expressed as the multiplication of the Gammatone filter ($G_T(f)$) and the symmetric filter ($e^{c\Theta(f)}$).

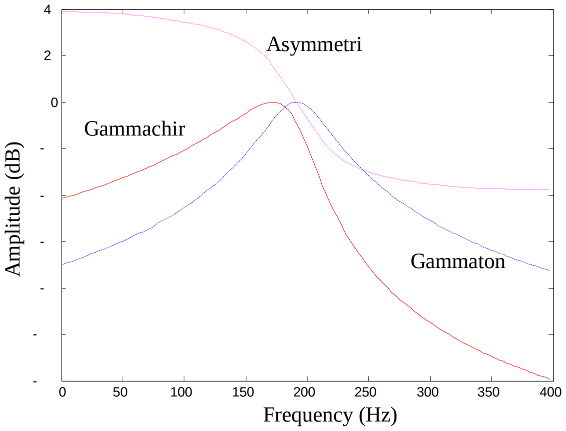

Figure 4. Gammatone function and asymmetric function synthesized Gammachirp function.

It could be clearly observed from Figure 5 that the frequency responses of Gammachirp filter with non-uniform bandwidths and significant asymmetry over the range of frequencies. The characteristics of the basement membrane could be illustrated.

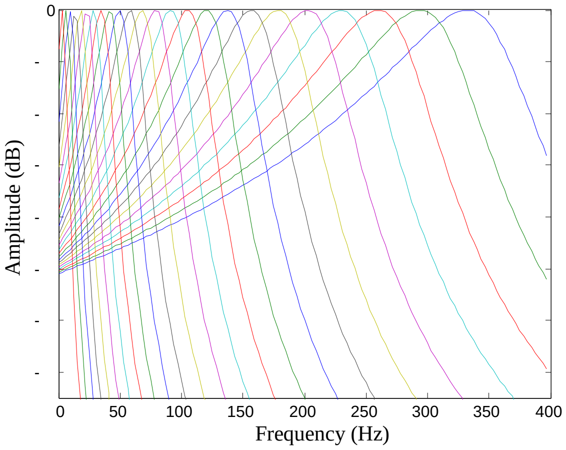

Figure 5. Example of 24 Gammachirp filterbank.

3.2. Feature Extraction with Gammachirp Filter Banks

The analysis of speech signals is processed by a Gammachirp filterbank. In this work, author used 24 Gammachirp filters in each filterbank, the filterbank is applied on the frequency band of $[0, fs/2]$ Hz, (where f_s is the sampling frequency, $fs = 8kHz$). The following three techniques are applied in the feature extraction: cube-root compression, RASTA, and CMVN. An illustrative block diagram is presented to demonstrate each step of the proposed feature extractor in Figure 6. Firstly, the cube-root compression is substituted with the output of each gammachirp filter such that the nonlinear of human auditory could be imitated.

$$Z_m = (Y_m)^{\frac{1}{3}} \quad (17)$$

Secondly, RASTA filter techniques a method to minimize the convolutional noise caused by the transmission channels.

$$R(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})} \quad (18)$$

Finally, the CMVN method is used to compensate for the effect of channel convolution noise in the cepstral domain.

Where y_m is output of the m^{th} filter.

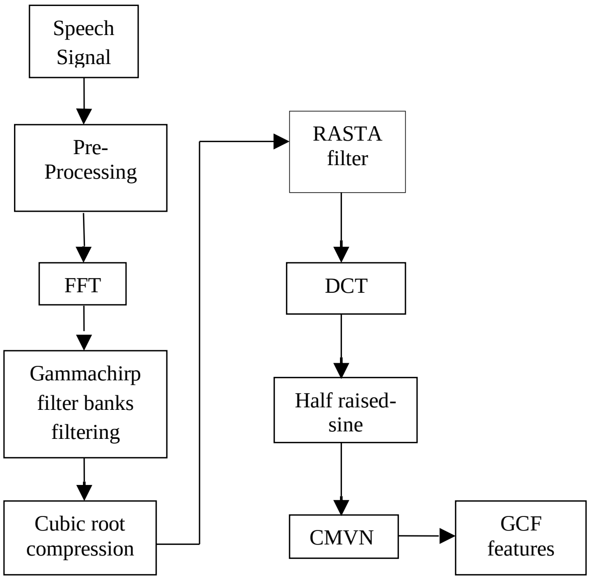

Figure 6. Block diagram of the used GCF method.

4. Experiment Setup

All the experimental data are using the Mandarin speech database collected by the laboratory where author studied in. Firstly, there are 120 speaker utterances (60 males and 60 females) as UBM training. Secondly, in addition to the above 120 speakers, the three experiments were carried out where there were total 36 speakers (22 males and 14 females) participating. Each speaker participated in two different recordings for the training and testing set. It should be noted that the training and testing set for each person is individual and has nothing to do with gender. In the training set, there are one utterance for each speaker and a total of 36 utterances. In addition, the average speech duration from the training data is around 60s for each speaker. Meanwhile, in the testing set, there are 5 utterances for each speaker and a total of 180 utterances. The duration of each testing utterance is approximately 5s of speech.

- *Experiment* 1: the performance of the proposed Gammachirp Cepstral Frequency (GCF) feature under different chip factor *c*.

The variation of the filter bandwidth and the spectral asymmetry property of the Gammachirp filter were considered in experiment 1. Besides, the level-dependent of signal will also be considered. In particular, if the level-dependent was considered in the experiment, author would manually adjust the chirp factor *c* to fit the asymmetric degree size of filter spectrum. In general, the range of the chirp factor *c* [10] is [-3, 3]. Therefore, the chirp factor in this experiment was chosen to be *c*=1, *c*=2, *c*=3 and *c*=-1, *c*=-2, *c*=-3 respectively. Followed by the above procedures, the performance of the proposed feature GCF could be tested. The mixed degree of GMM-UBM were 128, 256 and 512. The training data and testing data were both under clean testing condition.

- *Experiment* 2: Test the anti-noise capability for GCF based GMM-UBM.

The author uses NOISEUS database which includes three different types of ambient background noise: white noise, pink noise and f16 noise. The training data was under clean condition. And the testing data were collected by mixing clean utterances with four different noises at five different SNRs: 10, -5, 0, 5 and 10dB. During that process, Mel filter bank, Gammatone filter bank and Gammachirp filter bank were used respectively. In this experiment, 24 channels were set for each filterbank. Moreover, MFCC, RASTA-PLP, CFCC, GFCC, and GCF were chosen to be the feature extractors in this set of experiments. The chirp factor *c* of GCF is assumed to be 2. Lastly, GMM-UBM was used for the classifier whose mixed degree was 128.

- *Experiment* 3: Test the anti-noise capability of different chirp factor *c*.

The chirp factor in this experiment was chosen to be *c*=1, *c*=2, *c*=3 and *c*=-1, *c*=-2, *c*=-3 respectively. Followed by the above chirp factor *c*, the performance of the proposed feature GCF could be tested. GMM-UBM was used for the classifier whose mixed degree was 128.

5. Experiment Results

The result of experiment I are shown in Figure 7, 8, and 9. From Figure 9, it can be observed that the GCF feature generates the highest recognition rates while the mixture degrees of GMM-UBM were 128, 256 and 512. In addition, if the chirp factor *c* is positive, the difference of the recognition rate is always less than 1.11%. Eventually, the recognition rate can reach above 98%. However, if the chirp factor *c* is negative. When the mix degree of GMM-UBM is 128, the recognition rate of *c*=-3 is slightly less than that of 256 and 512. Furthermore, Figure 8 shows that there is not much difference of the recognition rates for c=-1, *c*=-2, c=-3. In conclusion, the experimental results below imply that the chirp factor *c* and the mix degree of GMM-UBM influence the recognition rates slightly.

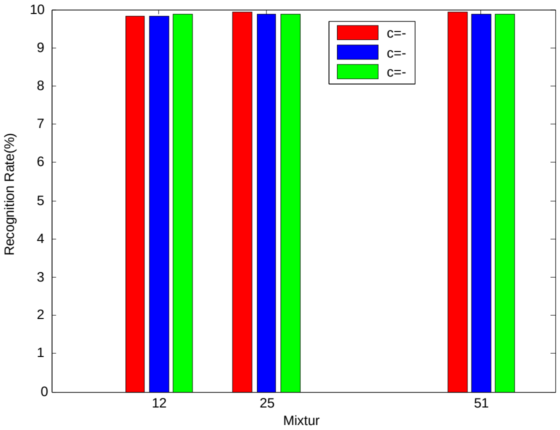

Figure 7. Recognition rate of different chip factor *c* (positive) under clean speech (%).

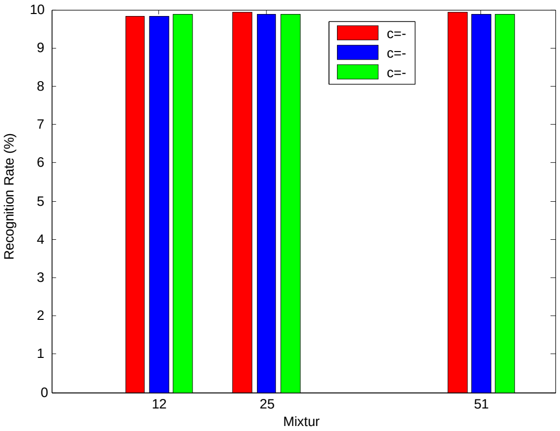

Figure 8. Performance comparison of different chip factor *c* (negative) under clean speech (%).

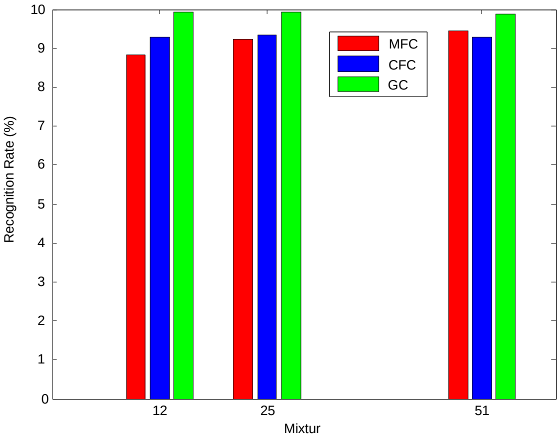

Figure 9. Performance comparison of MFCC, GFCC and proposed GCF (chirp factor *c* is 3) under clean condition (%).

The results of experiment 2 are summarized and shown from Figures 10, 11, and 12. These figures recorded the recognition rates obtained from the proposed GCF feature, GFCC feature and MFCC feature with three different types of ambient background noise at SNR of -10, -5, 0, 5 and 10dB respectively. In each figure, the effectiveness of proposed feature could be compared to other two features given the same SNR of a particular ambient background noise. For example, assuming that f16 noise is at 0dB SNR, the recognition rates of MFCC, RASTA-PLP, CFCC and GFCC features are 33.33%, 45.78%, 75.67% and 88.89% respectively, however, the GCF feature has the corresponding recognition rate of 93.89%. Compared to the MFCC feature, RASTA-PLP feature and CFCC feature, the GFCC feature has a higher recognition rate. However, it is still not as efficient as the proposed GCF feature. Moreover, it is clear to observe from these figures that when ambient background noise has the SNR higher than 5dB, the recognition rate of the proposed GCF feature is most surely higher than 93%.

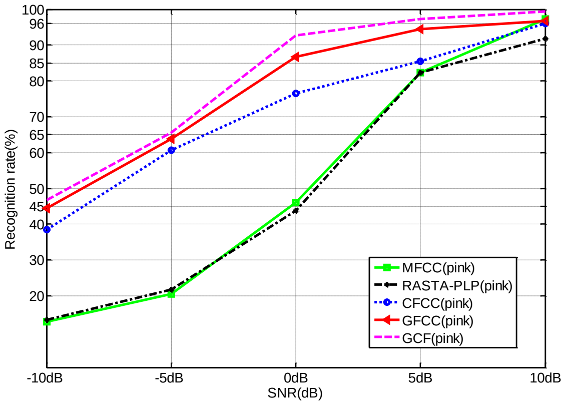

Figure 10. GMM-UBM (pink noise).

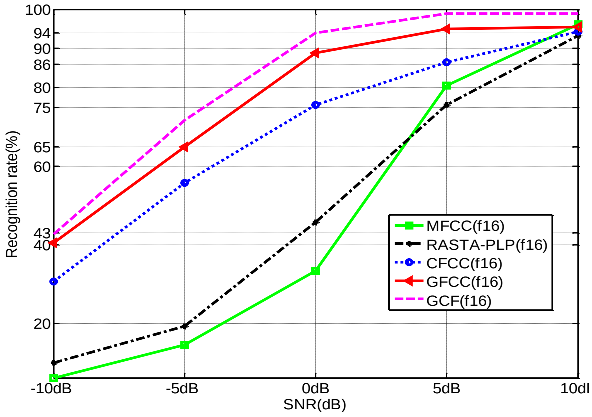

Figure 11. GMM-UBM (f16 noise).

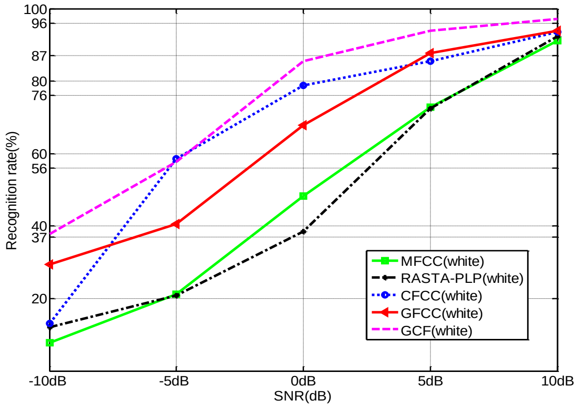

Figure 12. GMM-UBM (white noise).

The results of experiment 3 are summarized and shown from Table 1 to Table 3. These tables recorded the recognition rates obtained from the proposed GCF feature with three different types of ambient background noise at SNR of -10, -5, 0, 5 and 10dB respectively. In each table, the effectiveness of proposed feature could be compared with each other given the same SNR of a particular ambient background noise. For example, if the chirp factor *c* is positive and the SNR of the ambient background noise increases from -10dB to -5dB, the difference of the recognition rate is always less than 5%. Eventually, the recognition rate can reach above 90% with the SNR higher than 5dB. For example, if ambient background noise is pink noise and it is at the SNR of 10dB, the recognition rates with *c*=1, *c*=2, and *c*=3 are 97.78%, 99.44%, and 98.33% respectively. However, if the chirp factor *c* is negative and the ambient background noise is pink noise, the recognition rate of *c*=-3 is slightly less than that of *c*=-1, *c*=-2. Furthermore, Table 2 shows that there is not much difference of the recognition rates for *c*=-1, *c*=-2. In conclusion, the experimental results below imply that the chirp factor *c* could influence recognition rates.

Table 1. Speaker recognition rate with pink noise (%).

	SNR (10dB)	SNR (-5dB)	SNR (0dB)	SNR (5dB)	SNR (10dB)
c=-1	46.67	63.68	94.44	97.22	97.78
c=-2	46.67	68.33	93.89	97.22	97.78
c=-3	42.68	64.78	88.89	93.89	96.11
c=1	42.89	64.44	94.44	97.22	97.78
c=2	46.67	65.56	92.78	97.22	99.44
c=3	47.44	68.56	93.33	96.67	98.33

Table 2. Speaker recognition rate with f16 noise (%).

	SNR (10dB)	SNR (-5dB)	SNR (0dB)	SNR (5dB)	SNR (10dB)
c=-1	39.36	66.44	91.67	97.68	98.33
c=-2	39.22	66.00	91.68	97.68	98.78
c=-3	39.56	67.79	92.22	97.78	98.33
c=1	40.00	67.26	92.76	97.22	97.78
c=2	42.78	71.67	93.89	98.89	98.89
c=3	41.67	68.89	93.89	98.11	98.33

Table 3. Speaker recognition rate with white noise (%).

	SNR (10dB)	SNR (5dB)	SNR (0dB)	SNR (5dB)	SNR (10dB)
c=-1	36.11	54.56	82.33	98.33	97.22
c=-2	36.89	57.22	82.67	97.78	97.78
c=-3	40.56	58.56	85.56	93.89	97.78
c=1	35.63	53.69	84.22	91.67	97.89
c=2	37.78	57.78	85.56	93.89	97.22
c=3	40.49	58.22	86.11	93.44	97.78

6. Conclusions

In this study, authors proposed a robust feature extractor based on the Gammachirp filter and characteristics of human auditory system. And the authors also demonstrated that GMM-UBM can be applied for speaker recognition. The cube-root compression method, RASTA and CMVN were applied to the robust feature extraction. Two experiments were carried out and the results showed that the proposed feature extractor outperformed all the other feature extractors both under clean testing condition and noise testing condition. Moreover, under clean condition, the highest recognition rates were observed when the chirp factor *c*=3, which suggests that the choice of chirp factor *c* could significantly influence the recognition rate. Nevertheless, the chirp factor *c* is not the only factor that could affect the recognition rate, and the number of channels of the filter could also affect the recognition rate. Therefore, a promising direction for future work is to explore how the filter channels affects recognition rate.

References

[1] Abdallah A. and Hajaiej Z., "Improved Closed Set Text Independent Speaker Identification System Using Gammachirp Filterbank in Noisy Environments," *in Proceedings of 11th International Mulit-conference on Systems, Signals and Devices*, Barcelona, pp. 1-5, 2014.

[2] Abushariah M., Ainon R., Zainuddin R., Elshafei M., and Khalifa O., "Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on A Phonetically Rich and Balanced Speech Corpus," *The International Arab Journal of Information Technology*, vol. 9, no. 1, pp. 84-93, 2012.

[3] Bouchamekh M., Bousseksouand B., and Berkani D., "Gammachirp Filterbank Based Speech Analysis for Speaker Identification," *in Proceedings of 8th International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics*, USA, pp. 19-23, 2009.

[4] Chougule S. and Chavan M., "Channel Robust MFCCs for Continuous Speech Speaker Recognition," *Advances in Signal Processing and Intelligent Recognition Systems*, vol. 264, pp. 557-568, 2014.

[5] Chowdhury F., Selouani S., and O'Shaughnessy D., "Distributed Automatic Text-Independent Speaker Identification Using GMM-UBM Speaker Models," *in Proceedings of Canadian Conference on Electrical and Computer Engineering*, St. John's, pp. 372-375, 2009.

[6] Chu K. and Leung S., "SNR-dependent Nonuniform Spectral Compression for Noisy Speech Recognition," *in Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Montreal, pp. I-973, 2004.

[7] Hermansky H. and Morgan N., "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[8] Hermansky H., Morgan N., Bayya A., and Kohn P., "RASTA-PLP Speech Analysis Technique," *in Proceedings of IEEE International Conference Acoustics Speech Signal Processing*, San Francisco, pp. 121-124, 2002.

[9] Irino T. and Patterson R., "A Time-Domain, Level-Dependent Auditory Filter: The Gammachirp," *The Journal of Acoustical Society of America*, vol. 101, no. 1, pp. 412-419, 1997.

[10] Irino T. and Unoki M., "A Time-Varying, Analysis/Synthesis Auditory Filterbank Using the Gammachirp," *in Proceeding of IEEE International Conference Acoustics Speech and Signal Processing*, Seattle, pp. 3653-3656, 2002.

[11] Jacobson G., "Magnetoencephalographic Studies of Auditory System Function," *Journal of Clinical Neurophysiology*, vol. 11, no. 3, pp. 343-364, 1994.

[12] Jin Q., Robust Speaker Recognition, Theses, Carnegie Mellin University, 2007.

[13] Li Q. and Huang Y., "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions," *IEEE Transactions on Audio Speech Language Processing*, vol. 19, no. 6, pp. 1791-1801, 2011.

[14] Li Q. and Huang Y., "Robust Speaker Identification Using an Auditory-Based Feature," *in Proceedings of IEEE International Conference on Acoustics*, Speech and Signal Processing, Dallas, pp. 4514-4517, 2010.

[15] Mammone R., Zhang X., and Ramachandran R., "Robust Speaker Recognition: A Feature-Based Approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58-71, 1996.

[16] Muda L., Begam M., and Elamvazuthi I., "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138-143, 2010.

[17] Prasad N. and Umesh S., "Improved Cepstral Mean and Variance Normalization Using Bayesian Framework," *in Proceeding of IEEE Automatic Speech Recognition and Understanding*, Olomouc, pp. 156-161, 2013.

[18] Reynolds D. and Rose R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[19] Salhi K., Hajaiej Z., and Ellouze N., "A Novel Approach for Auditory Spectrum Enhancement to Improve Speech Recognition's Robustness," *in Proceedings of 12th IEEE International Multi-Conference on Systems, Signals and Devices*, Mahdia, pp. 1-5, 2015.

[20] Tazi E., Benabbou A., and Harti M., "Efficient Text Independent Speaker Identification Based on GFCC and CMN Methods," *in Proceedings of International Conference on Multimedia Computing and Systems*, Tangier, pp. 90-95, 2012.

[21] Vikram C. and Umarani K., "Text Independent Classification of Normal and Pathological Voices Using Mfccs and GMM-UBM," *in Proceedings of IEEE International Conference on Information and Communication Technologies*, Thuckalay, pp. 980-985, 2013.

Lei Deng was born in Sichuan, China in 1993. She received the B.S. degree from the College of information science and technology, Chengdu University of technology, Chengdu, China in 2015.She is currently pursuing the M.S. degree at the College of Electronics and Information Engineering, Sichuan University, Chengdu, China. Her research area manly includes speaker recognition, language identification, and speech signal processing.

Yong Gao (Corresponding author: gaoyong@scu.edu.cn) was born in Xi'an, China in 1969. He received the M.S. and Ph.D. degrees from the school of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, in 1997 and 2000, respectively. He is a professor in College of Electronics and Information Engineering, Sichuan University. His research area mainly includes speech signal processing, anti-interference and anti-interception technology in communication, modulation recognition, emergency communication, array signal processing, blind analysis of signal.