Sensors and Materials, Vol. 34, No. 7 (2022) 2809–2819
MYU Tokyo

S & M 3007

Recognition System for Cantonese Speakers in Different Noisy Environments Based on Estimate–Maximize Algorithm

Yu Fan,[1*] Chin-Ta Chen,[1] and Cheng-Fu Yang[2,3**]

[1]School of Electronic and Electrical Engineering College, Zhaoqing University, Zhaoqing City, Guangdong Province 526061, China
[2]Department of Chemical and Materials Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[3]Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung 413, Taiwan

(Received March 29, 2022; accepted May 24, 2022)

Keywords: speaker identification, estimate–maximize algorithm, Gaussian mixture model, Mel-frequency cepstrum coefficient, maximum likelihood estimation

Highly accurate personal identification systems are required in many different recognition situations. In this study, the Mel-frequency cepstrum coefficient was used to extract the features of speakers. The aim of this study was to identify different speeches in different noisy environments. A maximum likelihood estimation method based on noise probability was proposed to enhance the recognition effects of the Gaussian mixture model of different speeches from mixed noise speech signals. Experimental results indicated that the method had high recognition results under various noise conditions. The recognition results of the proposed method in different noise environments were superior to those of a method using only one type of noise for modeling. Experimental results obtained from some unspecified speakers showed that three different languages (Mandarin, English, and Cantonese) were effectively identified.

1. Introduction

Language is the most direct and effective communication method and tool for human beings because it is efficient, accurate, convenient, and natural. With the rapid development of society, many types of machines now participate in human production and social activities; therefore, improving the communication between people and machines and enhancing people's ease with operating machines have become increasingly important. Each person has distinctive phonetic characteristics that are difficult to imitate by other people. The acquisition of voice signals is convenient and simple, and the cost of voice capture equipment or systems is low. Voices not only have the characteristics of portability, uniqueness, and memorability, but also allow the use of simple and noncontact equipment for data acquisition. Other biological characteristics do not have these characteristics, and speaker recognition is also known as voice print recognition.

*Corresponding author: e-mail: fy@zqu.edu.cn
**Corresponding author: e-mail: cfyang@nuk.edu.tw
https://doi.org/10.18494/SAM3921

ISSN 0914-4935
https://myukk.org/

The recognition of different speakers and languages is an important branch of speech processing. Speaker recognition encompasses two fundamental tasks: speaker verification and speaker identification. Speaker verification is the task of determining the contents of speech from a set of speakers or known voices. Speaker identification is the task of determining whether the person is who they claim to be. Thus, to identify unknown speakers who do not match any of the models during the training processes, the investigation of an additional decision method or technology is necessary.(1)

Speaker recognition technologies have shown promising application prospects and made major progress, but many problems remain to be solved. Currently, a quiet environment is usually necessary for speakers to train recognition systems. This is because in a noisy environment, the recognition performance of recognition systems markedly deteriorates, and recognition systems can exhibit better recognition performance in a quiet environment. For example, Wu and Wang proposed a robust feature parameter whose recognition accuracy reached 95% in a quiet environment, but when a noise with a signal-to-noise ratio (SNR) of 18 dB was added, the recognition accuracy dropped to about 60%.(2) Robust endpoint detection is one of the most important areas of speech recognition processing, and feature parameters obtained from endpoint detection have very high sensitivity to the training environment. An effective endpoint detection algorithm can not only reduce the noise obtained from silent segmentation, but also reduce the processing time. In the past, a noise-free environment has usually been necessary to process the traditional endpoint detection algorithm, resulting in processing for voice recognition having low robustness to noise.(3) Thus, an urgent task is to find a feature parameter that is robust in noisy environments and can sufficiently specify the characteristics of speech.

Cantonese is a language spoken by a large number of people, mainly in Guangdong Province, Macao, Hong Kong, Hainan, and Guangxi in southern China and also in overseas Chinese communities in Southeast Asia, North America, Singapore, Australia, and elsewhere. The first novelty of this study was that an algorithm for recognizing Cantonese was investigated. Among the research on speech recognition technologies, the development of speaker recognition systems having strong robustness to noise has played an important role. We propose a new algorithm that is robust in a noisy environment, for example, a voice mixed with the sound of flowing water or a thunderstorm. Another novelty of this study is that we evaluated the voice recognition rates of a speaker recognition system under noiseless and noisy environments with different types of noise and with three different languages (Mandarin, English, and Cantonese). Simulation and experimental results verified that the proposed algorithm has good recognition performances for the three languages in the presence of differently impulsive noises.

2. Speaker Recognition System

A flow chart of the investigated speaker recognition system is shown in Fig. 1. The input speeches were sampled and converted to digital signals. The digital signals were input in the system and were subjected to classification and training phases. In the classification phase, the input speech was compared with stored reference models to identify the speech. In the training

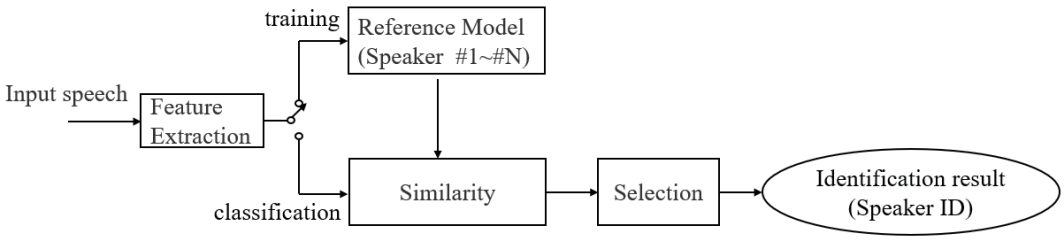

Fig. 1. Flow chart of speaker recognition system.

phase, reference models were trained for each registered speaker who provided samples of their speeches.

2.1 Feature extraction

The features of digital signals were extracted to convert the waveforms of the speech signals to a set of features for further recognition processes. Mel-frequency cepstrum coefficients (MFCCs) are defined as the coefficients collected from a Mel-frequency cepstrum (MFC). The feature vectors were extracted from the digital signals of the input speech in the format of MFCCs.[4,5] MFCCs were chosen because they are based on the perceptual characteristics of the human auditory system.[6,7] A block diagram of the MFCC feature extraction process is shown in Fig. 2. In the first step, speeches are sampled and converted to digital format, and they are used as input signals. Then, a Hamming window is applied to the digital speech signals, and this process divides the input signals into frames of samples. Each frame is windowed to minimize the discontinuity of the signal at the beginning and end of each frame. The fast Fourier transform (FFT) is used to convert each frame of samples from the time domain into the frequency domain. Then, filter banks are used to convert the frequency scale from hertz to the Mel-scale, with the frequency spaced linearly at lower frequencies and logarithmically at higher frequencies, then the logarithm is taken. The purpose of this stage is to capture the phonetically important characteristics of speech in a manner that reflects the human perceptual system.

2.2 Gaussian mixture model

A Gaussian mixture model (GMM) is a probabilistic model, and within an overall population, it can represent normally distributed subpopulations. GMM-based classifiers show good performance in many applications, including speech processing applications.[8–11] Therefore, GMMs are used as the basis of both the classification and training processes to classify the speakers based on the probability that the test data originated from each speaker in the set.[12]

2.2.1 Training phase

A GMM is essentially a multidimensional probability function. In the training phase, the GMM is established for each target speaker's speech. Several Gaussian functions are used as a

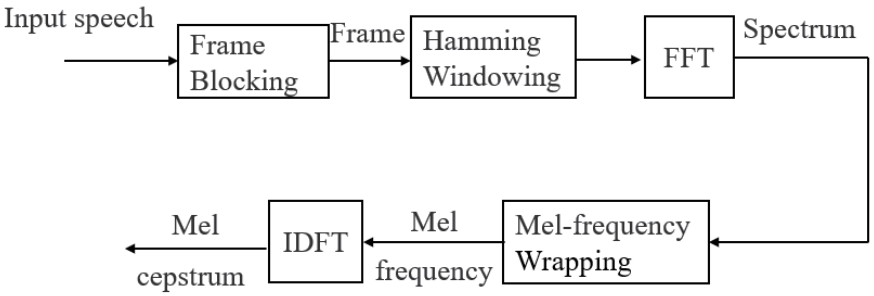

Fig. 2. Block diagram of MFCC feature extraction.

linear weighting to fit the probability distribution of the target speaker's speech feature vector o. For each speaker, a statistical model in the set is denoted by λ. For instance, the set of speaker s with size S is written as

$$\lambda_s = \{w_i, \mu_i, \sigma_i\} \quad i = 1,2,\cdots,M; \;\; s = 1,2,\cdots,S, \tag{1}$$

where, w, μ, and σ are the weight, mean, and diagonal covariance, respectively, and M is the number of GMM components. If the parameter set of a GMvM value with mixing degree M is λ, then λ can be used to represent a GMM value, and a linear combination of P single Gaussian distributions can be used to describe the Mth-order GMM values as follows:

$$P(o \mid \lambda) = \sum_{i=1}^{M} P(o, i \mid \lambda) = \sum_{i=1}^{M} w_i P(o \mid i, \lambda), \tag{2}$$

where w_i and o are the mixed weights of the ith component and the F-dimensional acoustic characteristic vector, respectively, and w_i is the prior probability of the given component. Note that

$$\sum_{i=1}^{M} w_i = 1. \tag{3}$$

$P(o \mid i, \lambda)$ is the ith GMM component; thus,

$$P(o \mid i, \lambda) = N(o \mid \mu_i, \Sigma_c) = \frac{1}{(2\pi)^{\frac{F}{2}} |\Sigma_c|^{\frac{1}{2}}} \exp\left\{ -\frac{(o-\mu_i)^T \sum_i^{-1} (o-\mu_i)}{2} \right\}. \tag{4}$$

Establishing a GMM value for speakers involves estimating the parameters of GMM values through processing the data of the target speakers. Maximum likelihood estimation (MLE) is a common method used to estimate the parameters.[13,14] We use the estimate–maximize (EM)

algorithm to find the MLE values from the GMM value. The training phase consists of two steps, initialization and the EM algorithm. The initialization step provides the initial estimations of the mean values for each Gaussian component in a GMM value. The EM algorithm recomputes the weights, covariances, and means of each component in a GMM value. Each iteration of the algorithm provides more accurate estimations of all three parameters. Two steps are performed to find the best parameters:

1. E-step:

x is assumed to be observable and generated by a probability distribution, and *x* is referred to as incomplete data. We denote the complete data set as $z = (x, i)$ and assume that the joint density function is

$$p(z \mid \lambda) = p(x, i \mid \lambda) = p(i \mid x, \lambda)\, p(x \mid \lambda). \tag{5}$$

Then, an auxiliary function is defined as

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{M}\sum_{t-1}^{T} \frac{w_i P(o_t \mid c, \lambda)}{P(o_t \mid \lambda)} \log \bar{w}_i + \sum_{i=1}^{M}\sum_{t=1}^{T} \frac{w_i P(o_t \mid c, \lambda)}{P(o_t \mid \lambda)} \log P(o_t \mid c, \bar{\lambda}). \tag{6}$$

2. M-step:

To estimate $\bar{w}_i, \bar{\mu}_i, \text{and}\, \bar{\sigma}_i$, the derivatives of the auxiliary function are taken with respect to $\bar{w}_i, \bar{\mu}_i, \text{and}\, \bar{\sigma}_i$ to estimate the *i*th weight:

$$\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} \frac{w_i P(o_t \mid c, \lambda)}{P(o_t \mid \lambda)} = \frac{1}{T}\sum_{t=1}^{T} P(c \mid o_t, \lambda). \tag{7}$$

The new estimations of the mean values can be written as

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} P(c \mid o_t, \lambda) o_t}{\sum_{t=1}^{T} P(c \mid o_t)}, \tag{8}$$

and the new estimations of the diagonal elements of the *i*th covariance matrix can be written as

$$\bar{\sigma}_{if}^2 = \frac{\sum_{t=1}^{T} P(c \mid o_t, \lambda)\left(o_{tf} - \mu_{tf}\right)^2}{\sum_{t=1}^{T} P(c \mid o_t, \lambda)}, \quad f = 1, 2, \cdots, F. \tag{9}$$

2.2.2 Classification

To classify the input digital signals, we assume that there are *N* target speakers, each of whom is represented by a GMM value, denoted by $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$, and $\{o_1, o_2, \cdots, o_T\}$ is used to recognize the observed feature sequence of speech. In this speaker recognition system, it is

necessary to identify the contents of speakers from speeches. This means that when the posterior probability of O for each GMM value is calculated, the speakers with the maximum posterior probability are determined to be those to which speech O belongs. Therefore, for each s in the speaker set, the classification system is used to find S in Eq. (1), and $p(\lambda_n \mid O)$ is maximized using the equation

$$p(\lambda_n \mid O) = \frac{p(O \mid \lambda_n) p(\lambda_n)}{p(O)}. \tag{10}$$

The recognition result is defined by the maximum posterior probability criterion and expressed as

$$n^* = \arg\max_{1 \le n \le N} p(\lambda_n \mid O). \tag{11}$$

3. Experimental Results and Discussion

3.1 Speech model

The language database used as the references was recorded in an interference-free and quiet environment. The sampling frequency used was 11.025 kHz with mono recording and 16-bit quantization. The entire database included 50 Cantonese speakers, and all speakers spoke standard Cantonese. Speech acquisition was divided into test speech and training speech, and the speakers read Chinese text aloud using normal speech. After the voices were recorded, they were stored in the form of a .wav file, and a specified folder was used to save them. The training time of each speech was 30 s and the test time of each speech was 10 s. In this experiment, we first used different MFCC dimension numbers to test the recognition rate of clean speech, then we added a single background noise. Finally, we tested the recognition rates of three different languages (Mandarin, English, and Cantonese) in a noiseless environment.

3.2 Clean voice waveform and spectrum

Figure 3 shows that the clean voice waveform and spectrum are distributed in the low-frequency region. In Fig. 4, the recognition rates with different Gaussian and MFCC dimension numbers are compared. When the Gaussian number was 2, the recognition rate had the lowest value but the recognition time was also shortest. The recognition rate increased from 82 to 91% and the recognition time fluctuated around 1.0 ± 0.05 s as the MFCC dimension number increased from 2 to 12. When the Gaussian number was 4, the recognition rate increased from 86 to 96% and the recognition time fluctuated around 2.2 ± 0.1 s as the MFCC dimension number increased from 2 to 12. When the Gaussian number was 8, the recognition rate increased from 93% to close to 100% and the recognition time fluctuated around 4.5 ± 0.15 s as the MFCC dimension number increased from 2 to 12. To obtain a high recognition rate, the Gaussian number and MFCC dimension number should be set to 8 and 12, respectively.

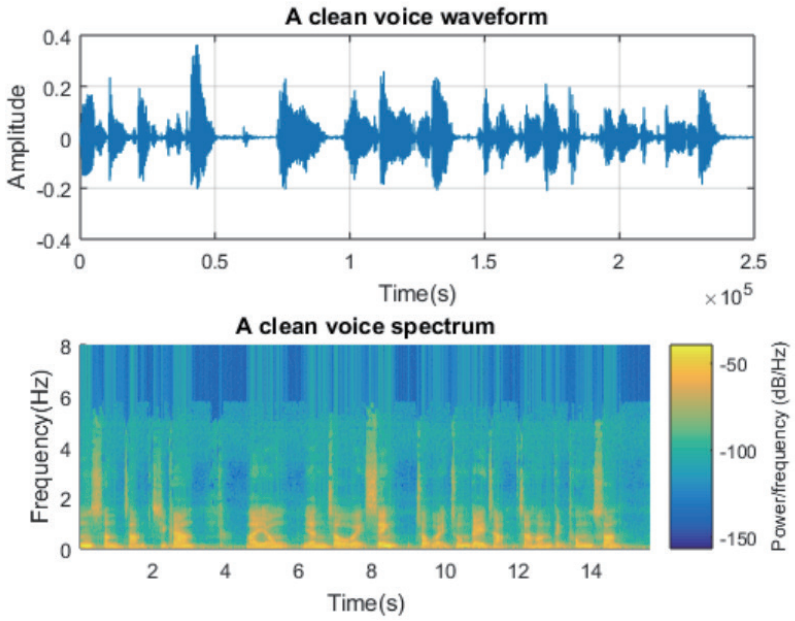

Fig. 3. (Color online) Clean voice waveform and spectrum.

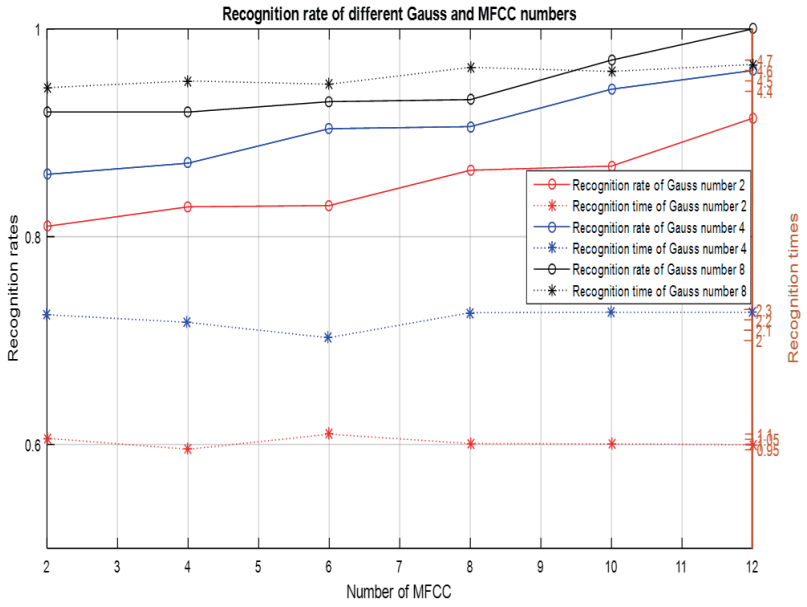

Fig. 4. (Color online) Curves of recognition rates and recognition times of clean voice with different Gaussian and MFCC dimension numbers.

3.3 Mixed noise waveforms and spectra

In a practical environment, a speech signal is often disturbed by different types of background noise, which cannot be controlled and can seriously affect the performance of an ideal recognition system. Usually, the background noise is an additive signal, and the collected signals

are the sum of a real speech signal and background noises. In this study, three types of noise signals, a thunderstorm, the noise of flowing water, and a thunderstorm together with the noise of flowing water, were used in the experiment. The noise signals with SNRs of 10 dB were incorporated into the recorded speech signals to generate noisy speech signals. Figures 5(a) and 5(b) display the waveforms and spectra of the mixed signals with the flowing water and thunderstorm noises. The waveform distribution of both noises in Fig. 5(a) was centered around

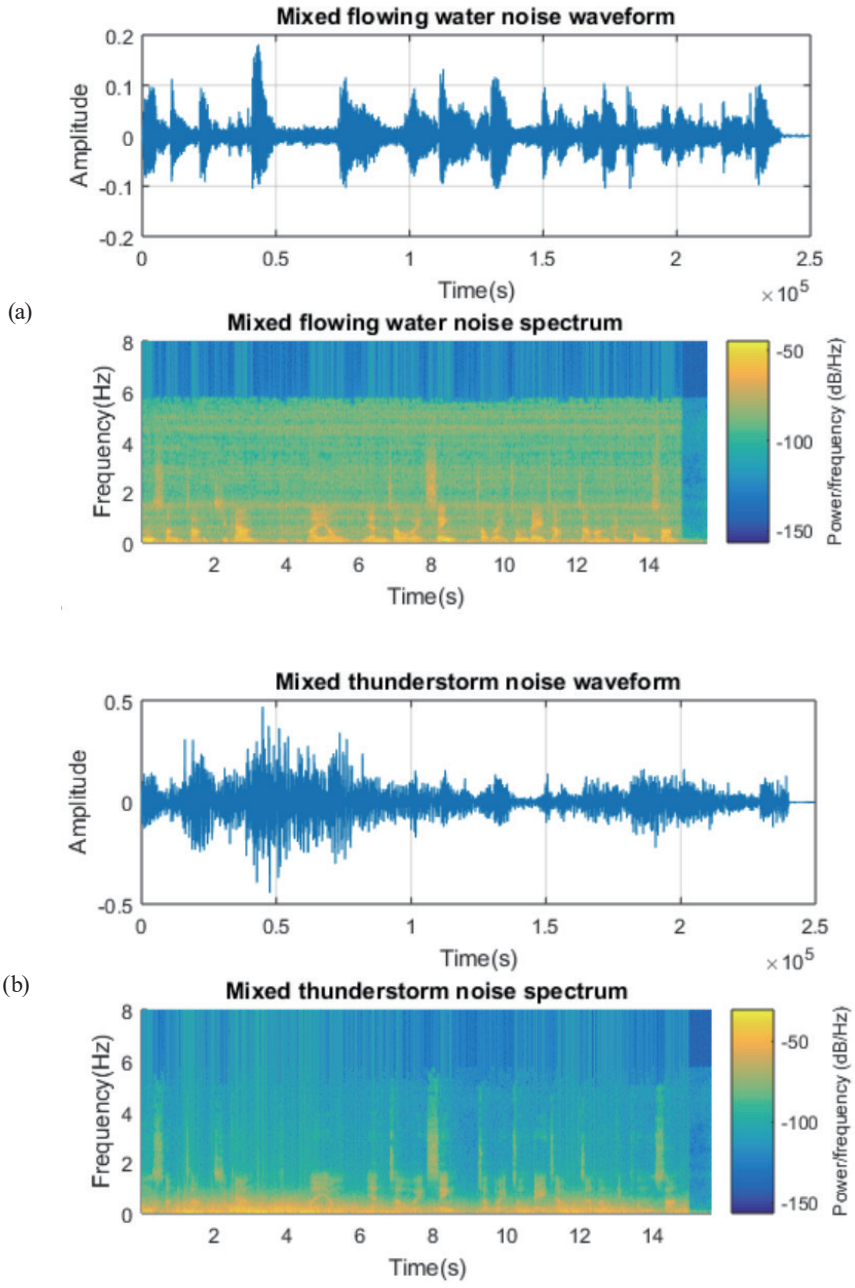

Fig. 5. (Color online) Mixed different noise waveforms and spectra. (a) Mixed with flowing water noise. (b) Mixed with thunderstorm noise.

0. The noise spectra were distributed in the high-frequency band: at around 3–6 Hz in Fig. 5(a) and 4–8 Hz in Fig. 5(b). Comparison of the results in Figs. 3 and 5 shows that the waveforms and spectra of the voice after mixing with the noises are markedly different from those of the clean voice.

As can be seen from Fig. 6, the recognition rate for the signal with the flowing water noise increased from 68 to 82% as the MFCC dimension number increased from 2 to 12 and that for the signal mixed with the thunderstorm noise increased from 64 to 78%. The recognition rate for the signal with both the flowing water and thunderstorm noises increased from 58 to 72% as the MFCC dimension number increased from 2 to 12. The recognition rate in the single-noise environment was lower than that in the clean speech environment, and the recognition rate was further reduced when both noises were mixed with the speech. Also, the lower the MFCC dimension number, the lower the recognition rate. These results show that the feature extraction of the speech signal is directly related to the MFCC dimension number, and the more accurate the feature extraction is, the higher the speech recognition rate will be.

3.4 Recognition rates for three different languages

The GMM speaker recognition model is currently the basic speaker recognition model, and the MFCC dimension number is the most widely used feature parameter in speech recognition, which can improve the recognition performance of a system, especially in an environment without noise. Therefore, we collected 50 samples of Mandarin speech and 50 samples of British English speech for speech recognition. As can be seen from the three broken lines in Fig. 7, the

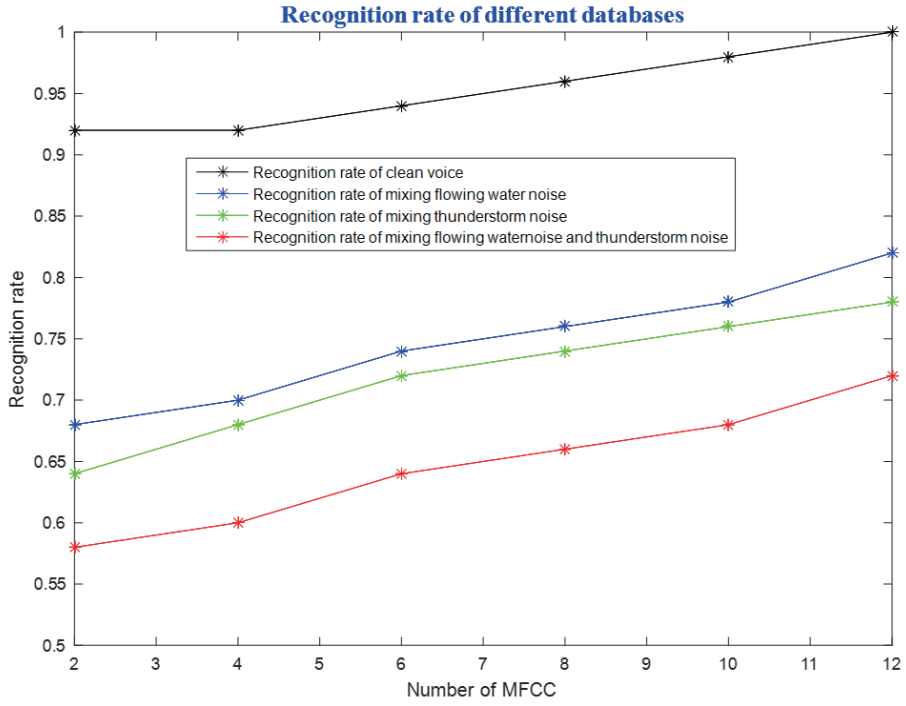

Fig. 6. (Color online) Curves of recognition rates with different noise conditions and MFCC dimension numbers.

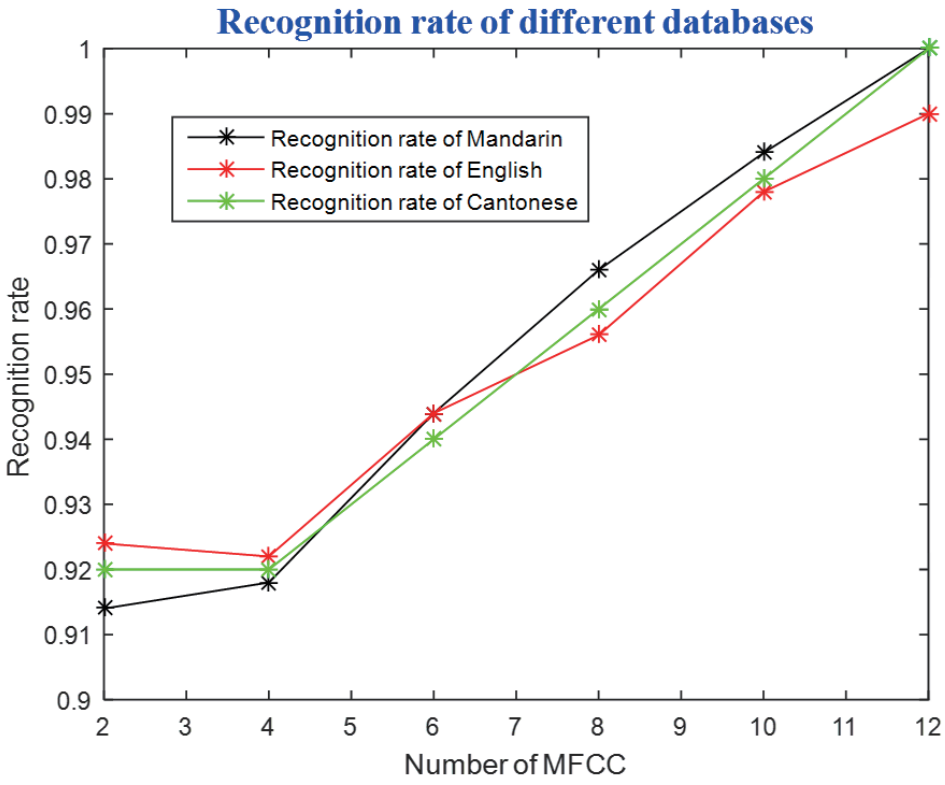

Fig. 7. (Color online) Curves of recognition rates for different languages.

recognition rate of Mandarin was slightly lower than those of English and Cantonese when the MFCC dimension number was 4 or less. When the MFCC dimension number was greater than 4, the recognition rate of Mandarin was higher than those of English and Cantonese; when the MFCC dimension number was 12, the recognition rates of Mandarin and English reached about 100%. It can be concluded that the speaker recognition system designed in this experiment is very effective for a small database of speaker speech.

4. Conclusions

In this study, we proposed a speaker recognition system to enhance the recognition rates of different languages under noiseless and different noise environments. To obtain a high recognition rate, the Gaussian number and MFCC dimension number should be set to 8 and 12, respectively. For different Gaussian numbers, the recognition rates of the clean voice increased and the recognition time showed no apparent fluctuation as the MFCC dimension number increased from 2 to 12. The recognition rates for the signals mixed with different noises (flowing water and/or thunderstorm noise) increased with the MFCC dimension number. The recognition rate of Mandarin was slightly lower than those of English and Cantonese when the MFCC dimension number was less than 4, and when the MFCC dimension number was 12, the recognition rates of Mandarin and English reached about 100%. In the future, once the

recognition effect of the investigated software has been further strengthened, we will import it into the hardware circuit design of a field-programmable gate array and realize applications to recognize different languages under a greater range of noise conditions.

Acknowledgments

This work was supported by projects under Nos. MOST 110–2622-E-390–002 and MOST 110–2221-E-390–020.

References

1 D. A. Reynolds and R. C. Rose: IEEE Trans. Speech Audio Process. **3** (1995) 72.
2 B. F. Wu and K. C. Wang: IEEE Trans. Speech Audio Process. **13** (2005) 762.
3 An Improved Endpoint Detection Algorithm Based on Improved Spectral Subtraction with Multi-taper Spectrum and Energy-Zero Ratio, T. Bao, Y. Li, K. Xu, Y. Wang, and W. Hu, Eds. (Springer International Publishing, New York, 2019) 1st ed.
4 Speech Synthesis and Recognition, J. N. Holmes and W. Holmes, Eds. (CRC Press, London, 2001) 2nd ed.
5 I. D. G. Y. A. Wibawa and I. D. M. B. A. Darmawan: J. Physics: Conf. Series **1722** (2021) 012014
6 R. Vergin, D. O'Shaughnessy, and A. Farhat: IEEE Trans. Speech Audio Process.**7** (1999) 525.
7 M. Hassan, M. Jamil, M. Rabbani, and M. Rahman: Proc. 3rd Int. Conf. Electrical & Computer Engineering (2004) 565–568.
8 M. Shi and A. Bermak: IEEE Trans. Very Large Scale Integr. VLSI Syst. **14** (2006) 962.
9 I. J. Ding and C. T. Yen: Multimed. Tools. Appl. **74** (2015) 5131.
10 Q. Feng, G. Hu, and X. Yao: Int. J. Electron. Commun. **62** (2008) 557.
11 V. Chauhan, S. Dwivedi, P. Karale, and S. M. Potdar: Int. Res. J. Eng. Technol. **3** (2016) 160.
12 Y. Fan, C. T. Chen, and C. C. Yang: 2021 4th Int. Conf. Data Science and Information Technology (2021) 49–54.
13 J. Xu, J. He, Y. Zhang, F. Xu, and F. Cai: Int. J. Distrib. Sens. Netw. **2016** (2016) 2080536.
14 M. S. Prasad and T. Panigrahi: Wireless Pers. Commun. **105** (2019) 1527.