Speaker recognition based on GMM-UBM and PLP

Chuyu Zhang ,Mingtao Wang
Xiaobo Li ,Yueyang Hu

Abstract— **Since the human ear has a strong ability to distinguish sounds, the speech characteristics based on the human auditory characteristics can reflect the essence of sound well and have good robustness. Therefore, speech feature extraction methods based on auditory characteristics are increasingly valued in the field of speech recognition. Based on the PLP (Perceptual Linear Predictive) characteristic parameter of the auditory characteristic, feature parameters are extracted from the collected voice information of different speakers. In order to be able to recognize the speaker's function for any piece of speech input, a Gaussian mixture model-generic background model (GMM-UBM) is used. The GMM-UBM model is trained with the obtained PLP characteristic parameters. After the model training is completed, a new voice is input for speaker recognition. When the Gaussian mixture component parameter is set to 32, the recognition accuracy reaches 97.82%.**

I. INTRODUCTION

Speech is the basic way for humans to exchange information, and it is also one of the most important biological characteristics of human beings. Due to the difference between the innate physiological differences and acquired behaviors, each person's speech has a strong personality, which makes it possible for us to recognize the speaker through speech. Speaker recognition as an important branch of speech processing technology has been widely studied in the field of pattern recognition and artificial intelligence. Different from speech recognition, it pays attention to the recognition of semantics. Speaker recognition pays more attention to the uniqueness of each speaker. The research focuses on the selection of feature parameters, the combination with classification models and the influence of noise. In terms of the selection of feature parameters, this paper selects the PLP characteristic parameters. The basis for the PLP characteristic parameters that can be used for speaker recognition is that the human ear has a strong ability to distinguish speech information from different speakers, and the PLP parameters are based on the human auditory model and are the result of the human auditory system. Acoustic features derived from the push. The PLP feature parameter extraction processes the speech signal in three aspects. The first is the critical band analysis process. A study of the human auditory mechanism found that when two tones with similar frequencies are simultaneously emitted, one can only hear one tone. The critical bandwidth refers to a bandwidth boundary where the subjective perception is abrupt. When the frequency difference between the two tones is less than the critical bandwidth, the person will listen to the two tones, which is called the shielding effect. The second point is that the equal-tone curve is pre-emphasized, and the equal-loudness curve is pre-emphasized with the equal-tone curve of the simulated human ear of about 40 dB. The third point is the signal strength-auditory loudness transformation, which aims to approximate the nonlinear relationship between the intensity of the simulated sound and the loudness of the human ear.

After obtaining the PLP characteristic parameters, in order to enable the computer to recognize different speakers, we need to extract the target user's sound extraction features into one or more models and store them in the model library. When testing or actually using, it is actually extracting the features in the currently received speech, comparing it with the model in the model library, and finally determining who is the speaker of the current speech. Therefore, the key to the recognition performance is the ability to model and distinguish the identity information in speech, and at the same time, it has sufficient anti-interference ability and robustness for the remaining information irrelevant to the identity.

Select GMM-UBM (Gaussian Mixture Model - General Background Model). GMM-UBM is an improved method for GMM. Since we can't collect enough voice from the target users, we can change the way of thinking from other The place collects a large number of non-target users' voices. We mix these non-target user data (the voiceprint recognition field is called background data) to fully train a GMM. This GMM can be seen as a representation of speech, but because it is trained from a large number of mixed data, it does not have the ability to characterize specific identities. This model can be thought of as a priori model of a specific speaker model. GMM-UBM gives a good pre-estimation of the probability model of spatial distribution of speech features, allowing us to train GMM in advance. Then the target user's data can be fine-tuned on this model. The most important advantage of the GMM-UBM model is that the model parameters are estimated by the MAP algorithm, which avoids the occurrence of over-fitting. At the same time, we do not have to adjust all the parameters (weight, mean, variance) of the target user GMM only for each Gaussian component. Mean parameters are estimated to achieve the best recognition performance. According to experiments, this can reduce the parameters to be estimated by more than half. The fewer parameters also mean faster convergence, and the model is well trained without so much target user data.

II. METHODS

A. PLP

Perceptual linear prediction (PLP) is a feature parameter based on auditory model. PLP is similar to LPC in that it is obtained by predictive coefficients, but PLP is a set of coefficients that can be used to predict polynomials by using all-pole model. Different from LPC, PLP technology applies some conclusions from human ear auditory test to spectrum analysis, and replaces the time-domain signal in linear prediction analysis with the signal obtained from the analog human ear auditory model.

PLP technology mainly imitates the auditory perception mechanism of human ear at three levels:

(a) Critical-band spectral resolution

(b) The equal-loudness pre-emphasis

(c) The intensity-loudness power law of hearing

1) PLP feature extraction flow chart:

Fig.1 is the principle block diagram of PLP feature parameter extraction. It can be seen that all kinds of characteristics of human ear hearing are processed by engineering and simulated by simple model. The spectrum obtained after processing is more in line with the characteristics of human ear hearing, so that the feature parameters have better robustness.

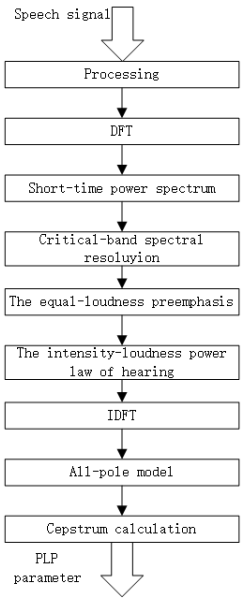

Fig. 1. PLP feature extraction

2) Spectrum analysis:

The preprocessing in flowchart means that the input speech signal is sampled and quantized first, and then framed by adding windows, that is to say, the signal is analyzed in time domain. In the experiment, Hamming window is used to obtain a smoother short-term spectrum than rectangular window. It is generally believed that the speech signal is stable between 10 and 30 ms, so we can roughly determine the window length.

After pretreatment, the signal is transformed into frequency domain by discrete Fourier transform. The square sum of the real and imaginary parts of the short-term speech spectrum is taken to obtain the short-term energy spectrum (power spectrum) $P(f)$ of the speech signal.

$$P(f) = Re[X(f)]^2 + Im[X(f)]^2 \tag{1}$$

3) Critical band analysis:

Critical frequency band analysis reflects the masking effect of human ear hearing, and is the embodiment of human ear auditory model. It converts the short-term energy spectrum of speech signal into Bark spectrum which accords with the auditory characteristics of human ear. The whole critical frequency band analysis process is shown in the Fig.2.

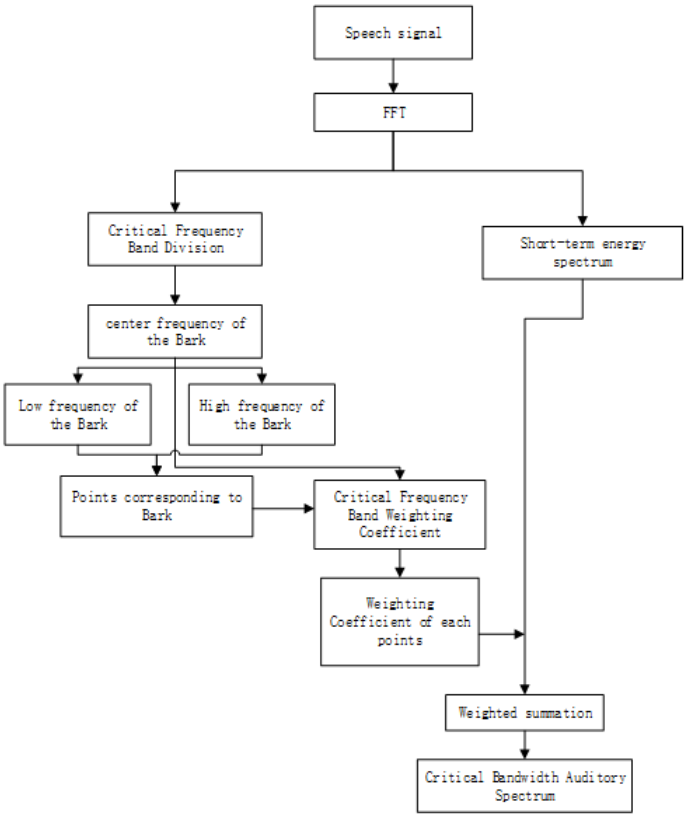

Fig. 2. Critical Frequency Analysis Flow

From the Eq.(2)

$$Z(f) = 6\ln\{f/600 + [(f/600)^2 + 1]^{0.5}\} \tag{2}$$

Mapping the frequency axis f of spectrum $P(f)$ to Bark frequency Z. After dividing the critical frequency band, the center frequency Z0(k) of each Bark band can be calculated. Generally, $Z_0(k) = 0.98kBark(k = 1, 2.\ldots\ldots)$.

According to $Z_0(k)$, the weighting coefficients of critical frequency band can be obtained by using the formula(3):

$$\Psi(Z-Z_0(k)) = \begin{cases} 0, Z - Z_0(k) < -1.3 \\ 10^{Z-Z_0(k)+0.5}, -1.3 \leq Z - Z_0(k) < -0.5 \\ 1, -0.5 \leq Z - Z_0(k) < 0.5 \\ 10^{-2.5(Z-Z_0(k)+0.5)}, 0.5 \leq Z - Z_0(k) \leq 2.5 \\ 0, Z - Z_0(k) > 2.5 \end{cases} \tag{3}$$

At the same time, the formula (4) can be used to calculate the corresponding frequency $f_0(K)$ of $Z_0(k)$ (in units of Hz), and $F_0(K)$ can be used to calculate the weight coefficients of the pre-emphasis of equal loudness.

$$f_0(k) = 1/2 * 600 * (e^{Z_0(k)/6} - e^{-(Z_0(k)-2.5)/6}) \tag{4}$$

The corresponding low-end frequency $f_l(k)$ and high-end frequency $f_h(k)$ in each Bark can be obtained by formula (5)(6).

$$f_l(k) = 1/2 * 600 * (e^{(Z_0(k)-2.5)/6} - e^{-(Z_0(k)-2.5)/6}) \tag{5}$$

$$f_h(k) = 1/2 * 600 * (e^{(Z_0(k)+1.3)/6} - e^{-(Z_0(k)+1.3)/6}) \tag{6}$$

According to $fl(k)$ and $f_h(k)$, the lowest and highest points corresponding to each Bark can be obtained, and then the specific weighting coefficients of each point can be known. The critical bandwidth auditory spectrum $\theta(k)$ can be obtained by summing the weighted short-term power spectrum of speech.

$$\theta(k) = \sum_{N=n_l(k)}^{n_h(k)} p(f(N))\Psi(Z(N) - Z_0(k)) \tag{7}$$

$Z_0(k)$ denotes the central frequency of the auditory spectrum in the k^th critical band, $n_l(k)$ denotes the low end of the auditory spectrum in the k^th critical band, and $n_h(k)$ denotes the high end of the auditory spectrum in the kth critical band.

4) Pre-accentuation of isoloudness curve:

Pre-accentuation of the isoloudness curve by using the simulated human ear isoloudness curve $E(f)$ of about $40dB$:

$$\Gamma(k) = E[f_0(k)]\theta(k) \tag{8}$$

$f_0(k)$ denotes the frequency corresponding to the central frequency of the auditory spectrum in the K critical band.

$E[f_0(k)]$ is the weight coefficient of equal loudness pre-emphasis, which approximately reflects the sensitivity of human ears to different frequencies.

$$E[f_0(k)] = \frac{(f_0^2(k) + 1.44 * 10^6) f_0^4(k)}{(f_0^2(k) + 1.6 * 10^5)^2 + (f_0^2(k) + 9.61 * 10^9)} \tag{9}$$

5) Intensity loudness conversion:

In order to approximate and simulate the non-linear relationship between the intensity of sound and the response of human ear, the amplitude of loudness is compressed after the pre-emphasis of the isoloudness curve.

$$\Phi(k) = \Gamma(k)^{0.33} \tag{10}$$

6) Solving Linear Prediction Coefficient by Total Pole Model:

After intensity loudness conversion, PLP feature parameters are extracted by discrete Fourier transform, and then a set of coefficients of all-pole model is obtained, that is, linear prediction analysis. Linear Prediction Analysis (LPA) is one of the most effective speech analysis techniques. It contains the basic concept that the present value of a speech sample can be approximated by weighted linear combination of the past value of Ruoqian speech. The weighted coefficients in linear combinations are called linear prediction coefficients. The process of linear prediction analysis is the process of calculating linear prediction coefficients. By minimizing the sum of squares of the difference between the actual speech sampling and the linear prediction sampling, the unique group prediction coefficient can be determined.

7) Cepstrum calculation:

Cepstrum features contain more information than other parameters and can better represent speech signals. The relationship between cepstrum characteristic $c(n)$ and linear prediction coefficient $alpha_i$ can be seen in the formula(10).

$$\begin{cases} c(1) = \alpha_1, \\ c(n) = \alpha_n + \sum_{i=1}^{n-1} (1 - i/n)\alpha_i c(n-i), 1 < n \leq p \\ c(n) = \sum_{i=1}^{p} (1 - i/n)\alpha_i c(n-i), n > p \end{cases} \tag{11}$$

In the Eq.(11), n is cepstrum order and P is linear prediction model order. According to the concept of homomorphic processing and the model of speech signal generation, the cepstrum $C(n)$ of speech signal is equal to the sum of cepstrum of excitation signal and the cepstrum of channel transmission function. The former is widely distributed and can be extended from low time domain to high time domain, while the latter is mainly distributed in low time domain. The speech information carried by speech signal is mainly reflected in the channel transmission function, so in speech recognition, the low-time domain cepstrum of speech signal is usually used to form cepstrum characteristic parameters C.

$$c = [c(1), c(2), ..., c(n)] \tag{12}$$

B. MFCC

The human ear has different perceptual abilities to speech at different frequencies. It is found that the perception ability is linear with frequency below $1000Hz$, and the perception ability is logarithmically related to frequency above $1000Hz$.

In order to simulate the perception characteristics of the human ear to different frequencies, the concept of Mel frequency is proposed. The meaning is: $1Mel$ is $1/1000$ of the degree of pitch perception of $1000Hz$, and the conversion formula between frequency f and Mel frequency B is:

$$Mel(f) = 2595 \lg(1 + f_{Hz}/700) \tag{13}$$

Where f is the frequency and the unit is Hz.

The Mel Frequency Cepstral Coefficient (MFCC) is proposed based on the above-mentioned Mel frequency concept. The proposed process and calculation process are shown in Fig.3.

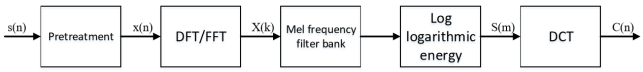

Fig. 3. the proposed process of MFCC

1) : The original signal s(n) is pre-emphasized, framed, windowed to the time domain signal $x(n)$ of each speech frame.

2) : After the time domain signal $x(n)$ is complemented by a number of 0, a sequence of length N (N should be an integer power of 2, where N is 256) is formed, and then a linear spectrum $X(k)$ is obtained by Discrete Fourier Transform, the conversion formula is:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} (0 \le n, k \le N-1) \tag{14}$$

In practical applications, it is often calculated by a Fast Fourier Transfor, where N is generally referred to as the window length of the FFT.

3) : The linear spectrum $X(k)$ is obtained by a Mel frequency filter bank to obtain a Mel frequency, and a logarithmic spectrum $S(m)$ is obtained by processing a logarithmic energy, where in the Mel frequency filter bank is set in a frequency spectrum range of the speech. A number of bandpass filters $H_m(k)$, $0 \le m \le M$, M is the number of filters, each filter has a triangular filter characteristic, and its center frequency is $f(m)$, when the m value is small The spacing between $f(m)$ is also small. As m increases, the interval between adjacent $f(m)$ also increases.

4) : Calculate the logarithmic energy of each filter bank output:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), (0 \le m < M) \tag{15}$$

5) : The above-mentioned logarithmic spectrum $S(m)$ is subjected to discrete cosine transform DCT to the cepstrum domain to obtain the Mel frequency cepstral coefficient (n):

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi n(m+1/2)}{M}\right), (0 \le m < M) \tag{16}$$

In the actual application process, the number of Mel filter banks is taken as 24, and the first 13 coefficients are selected.

C. GMM-UBM

GMM is a model composed of multiple Gaussian probability density functions, which is essentially a Likelihood Ratio Detector.

1) Likelihood Ratio Detector:

Given a segment of speech, Y , and a hypothesized speaker, S, the task of speaker detection, also referred to as verification, is to determine if Y was spoken by S. An implicit assumption often used is that Y contains speech from only one speaker.

The single-speaker detection task can be restated as a basic hypothesis test between H0(Y is from the hypothesized speaker S) and H_1(Y is not from the hypothesized speaker S).

The optimum test to decide between these two hypotheses is a likelihood ratio test given by:

$$\frac{p(Y \mid H_0)}{p(Y \mid H_0)} \begin{cases} > \theta, accept H_0 \\ < \theta, reject H_0 \end{cases} \tag{17}$$

where $p(Y \mid H_i), i = 0, 1$, is the probability density function for the hypothesis H_i evaluated for the observed speech segment Y , also referred to as the likelihood of the hypothesis H_i given the speech segment. 4 The decision threshold for accepting or rejecting H_0 is θ . The basic goal of a speaker detection system is to determine techniques to compute values for the two likelihoods, $p(Y \mid H_0)$ and $p(Y \mid H_1)$.

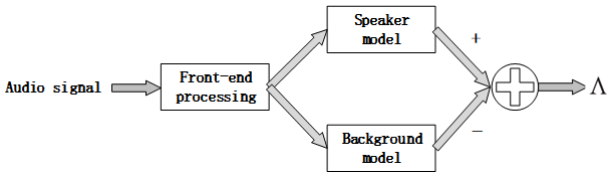

Fig. 4. Speaker detection systems based on likelihood ratios

Fig.4 shows the basic components found in speaker detection systems based on likelihood ratios. The role of the front-end processing is to extract from the speech signal features that convey speaker-dependent information. The output of this stage is typically a sequence of feature vectors representing the test segment, $= x_1, ..., x_T$, where x_t is a feature vector indexed at discrete time $t \in [1, 2, ..., T]$. There is no inherent constraint that features extracted at synchronous time instants be used; as an example, the overall speaking rate of an utterance could be invoked as a feature. These feature vectors are then used to compute the likelihoods of H_0 and H_1. Mathematically, H_0 is represented by a model denoted λ_{hyp} that characterizes the hypothesized speaker S in the feature space of x. For example, one could assume that a Gaussian distribution best represents the distribution of feature vectors for H_0 so that λ_{hyp} would be denoting

the mean vector and covariance matrix parameters of the Gaussian distribution. The alternative hypothesis, H_1, is represented by the model λ_{hyp}.

The likelihood ratio statistic is then:

$$p(X \mid \lambda_{hyp})/p(\overline{X} \mid \lambda_{\overline{hyp}}) \tag{18}$$

Often, the logarithm of this statistic is used giving the log-likelihood ratio

$$\Lambda(X) = log[p(X \mid \lambda_{hyp})] - log[p(\overline{X} \mid \lambda_{\overline{hyp}})] \tag{19}$$

While the model for H_0 is well defined and can be estimated using training speech from S, the model for $\lambda_{\overline{hyp}}$ is less well defined since it potentially must represent the entire space of possible alternatives to the hypothesized speaker. Two main approaches have been taken for this alternative hypothesis modeling. The first approach is to use a set of other speaker models to cover the space of the alternative hypothesis. In various contexts, this set of other speakers has been called likelihood ratio sets, cohorts, and background speakers. Given a set of N background speaker models $\{\lambda_1, ..., \lambda_N\}$, the alternative hypothesis model is represented by $p(X \mid \lambda_{hyp}) = F(p(X \mid \lambda_1), ..., p(X \mid \lambda_N))$, where $F()$ is some function, such as average or maximum, of the likelihood values from the background speaker set. In general, it has been found that to obtain the best performance with this approach requires the use of speaker-specific background speaker sets.This can be a drawback in applications using a large number of hypothesized speakers, each requiring their own background speaker set.The second major approach to alternative hypothesis modeling is to pool speech from several speakers and train a single model. Various terms for this single model are a general model , a world model, and a universal background model. Given a collection of speech samples from a large number of speakers representative of the population of speakers expected during recognition, a single model, λ_{bkg}, is trained to represent the alternative hypothesis. The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task. In this experiment we will use a single background model for all hypothesized speakers.

2) Gaussian Mixture Models:

An important step in the implementation of the above likelihood ratio detector is selection of the actual likelihood function, $p(X \mid \lambda)$. The choice of this function is largely dependent on the features being used as well as specifics of the application. For text-independent speaker recognition, where there is no prior knowledge of what the speaker will say, the most successful likelihood function has been Gaussian mixture models.

For a D-dimensional feature vector, x, the mixture density used for the likelihood function is defined as:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{20}$$

The density is a weighted linead combination of M unimodal Gaussian densities, $p_i(x)$, each parameterized by a mean $D * 1$ vector, μ_i , and a $D * D$ covariance matrix, $\sum_i$;

$$p_i(x) = \frac{1}{(2\pi)^{D/2}(\sum_i)^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)'(\sum\nolimits_i)^{-1}(x-\mu_i)\right\} \tag{21}$$

The mixture weights, w_i , furthermore satisfy the constraint $\sum_{i=i}^{M} w_i = 1$. Collectively, the parameters of the density model are denoted as $\lambda = \{w_i, \mu_i, \sum_i\}$, where $i = 1, \ldots, M$.

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation–maximization (EM) algorithm. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors, i.e., for iterations k and $k+1$, $p(X \mid \lambda(k+1)) > p(X \mid \lambda(k))$.

Usually, the feature vectors of X are assumed independent, so the log-likelihood of a model λ for a sequence of feature vectors, $X = \{x_1, \ldots, x_T\}$, is computed as

$$logp(X \mid \lambda) = \sum_{i=1}^{T} logp(x_i \mid \lambda) \tag{22}$$

Often, the average log-likelihood value is used by dividing $logp(X \mid \lambda)$ by T.

3) Universal Background Model(UBM):

In the GMM-UBM system we use a single, speaker-independent background model to represent $p(X \mid \lambda_{hyp})$. The UBM is a large GMM trained to represent the speaker-independent distribution of features. We simply merge all the training data to train a UBM.

III. Comparison and Analysis

A. Dataset description

We use the voice set of 30 people as a sample, each person's voice set consists of 290 short-term Chinese voices, a total of 8700 voices. Among them, there are 20 boys and 10 girls. Each speech is no more than 3 seconds. The detail is below:

TABLE I
DISTRIBUTION OF SPEAKERS

Speech Content	boys	girls	total
Train set	4060	2030	6090
Test set	1740	870	2610
Total	5800	2900	8700

B. Description of the experimental process

The speaker recognition experiment is divided into two stages: training stage and recognition stage. In training stage, train GMM-UBM model after extracting features. UBM is a higher-order GMM that takes many people's voices and trains them together. UBM is just trained once and can be

reused. In the training process, the GMM model of each speaker can be obtained through MAP self-adaptation. The operation process is as Fig5.

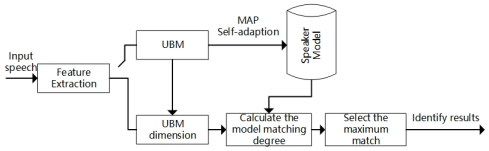

Fig. 5. Training Process

C. GUI

The GUI is shown in Fig.6.

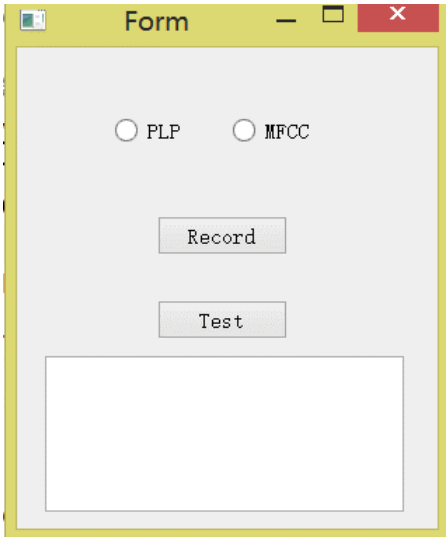

Fig. 6. GUI

The main operations are below:

(1) The Record button. It is used to record voice.

(2) The PLP and MFCC is used to select which feature parameter to extract. This is a single option.

(3) When we finish recording voice, click the Test button to calculate the probability of the identity of the speaker, which can call Baidu voice synthesis interface. The probability can be shown in the Text box.

D. Result analysis

We get the following experimental results by choosing different speech features and changing GMM parameter settings.

Among them, $N_components$ represent the number of Gauss functions, $Timespend$ is the time spent in recognition, $Trainacc$ is the training accuracy, and $Testacc$ is the testing accuracy.

The results show that MFCC and PLP have no significant effect on the recognition effect, but the number of Gauss function ($N_component$) has a significant effect on the testing accuracy. When $N_component = 1$, the testing speed is the fastest, but the recognition effect is intolerable. With the increase of $N_component$, the recognition time increases, the speed decreases, and the testing accuracy increases. When N is set to 8, the test accuracy reaches 95% and the test time is relatively short.

N_components	Timespend s	Trainacc %	Testacc %
1	125.06	3.07	4.02
2	145.18	76.64	74.63
3	214.49	89.48	89.55
4	186.45	92.32	91.92
5	206.87	93.57	93.03
6	235.93	94.62	93.72
8	278.89	95.65	95.10
10	400.18	96.13	95.37
12	411.26	96.95	95.83
14	448.88	97.16	96.10
16	529.22	97.29	96.33
24	909.91	98.08	96.94
32	1020.83	98.51	97.55
48	1152.34	99.06	98.01

Fig. 7. MFCC feature

N_components	Timespend s	Trainacc %	Testacc %
1	96.41	3.15	3.67
2	125.18	74.44	72.37
3	143.16	85.14	83.51
4	151.15	89.86	88.94
5	162.92	91.88	90.74
6	225.59	93.18	92.81
8	254.16	94.98	94.11
10	315.20	95.46	94.64
12	383.91	96.60	95.83
14	343.01	97.28	96.13
16	460.87	97.34	96.06
24	552.47	98.18	96.98
32	803.22	98.80	97.82
48	1188.11	99.05	98.16

Fig. 8. PLP feature

In addition, we do not use relevant texts in this experiment, which shows that it can achieve speaker recognition of non-relevant texts.

REFERENCE

[1]Yu-Min Zeng, Zhen-Yang Wu, Falk, T., Chan, W.-Y.. Robust GMM Based Gender Classification using Pitch and RASTA-PLP Parameters of Speech[P]. Machine Learning and Cybernetics, 2006 International Conference on,2006.

[2]Yucesoy, E.,Nabiyev, V.V.. Comparison of MFCC, LPCC and PLP features for the determination of a speaker's gender[P]. Signal Processing and Communications Applications Conference (SIU), 2014 22nd,2014.

[3]M.S. Athulya,P.S. Sathidevi. Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers[J]. Digital Investigation,2018,25.