International Journal of Control and Automation
Vol.8, No.3 (2015), pp.135-146
http://dx.doi.org/10.14257/ijca.2015.8.3.16

Speaker Recognition of Noisy Short Utterance Based on Speech Frame Quality Discrimination and Three-stage Classification Model

Ying Chen and Zhenmin Tang

School of Computer Science & Engineering, Nanjing University of Science and Technology, Jiangsu Nanjing 210094, China
chenying2124@163.com, tzm.cs@mail.njust.edu.cn

Abstract

The noisy short utterance is polluted by noise and corpus is less, so the recognition rate significantly decreased. For improving recognition rate, we proposed the dual information quality discrimination algorithm to classify the speech frames: one is differences detection and discrimination algorithm (DDADA), another is the improved SNR discrimination algorithm (ISNRDA). Based on the above two algorithms, the speech frames are classified to three classes: high quality, medium quality and low quality. We proposed GMM-UBM three-stage classification model, and we combine the dual information quality discrimination algorithm with GMM-UBM three-stage classification model. Experiments show that, the dual discrimination quality algorithms can be more precise to classify speech frame, and combining it with GMM-UBM three-stage classification model can make full use of limited corpus of short utterance and can improve the speaker recognition rate of the noisy short utterance.

Keywords: *noisy short utterance; noise separating; dual information quality discrimination algorithm; GMM-UBM three-stage classification model*

1. Introduction

After years of research, the field of speaker recognition has made a lot of research achievements [1-2]. At present, the rate of speaker recognition under laboratory environment (clean speech) has a very high. Mel frequency cepstral coefficients MFCC and universal background model UBM is widely used, document [3] extends the application of UBM model in noisy environment. Joint factor analysis (JFA) technology recently proposed by Kenny has opened up a new direction for the study of speaker recognition under channel mismatch, on the basis of this, Dehak proposed the concept of I- vector (Identity Vector, I-Vector) [4], the speech differences is represented in a low dimensional space. Vogt applied JFA and I- vector technology to speaker recognition of short utterance, based on Kenny [5].

In the real environment, separation of clean speech and background noise and accurate judgment of speech frames based on their quality, are helpful to improve the speaker recognition rate of noisy short utterance.

Document [6] uses all kinds of methods to get noise spectral and pure speech spectral, and classify them, and then compare these methods. All of these methods first estimated noise spectral, and then estimated pure speech spectral based on spectral subtraction, finally classified speech frames. But if the noise is nonstationary, the recognition rate of these methods is bad.

ISSN: 2005-4297 IJCA

In this paper, we use the dual information discrimination algorithm proposed to classify the speech frames, and combine the classification results with GMM-UBM three-stage classification model.

2. Speech Frames Quality Discrimination

2.1. Extract Spectral Feature of Noisy Speech

Spectral features are computed using the short-time Fourier transform (STFT) on a frame-by-frame basis. The signal is transformed into overlapping segments and the N-point STFT is computed as

$$X(i,k)=\sum_{n=0}^{N-1}w(n)x(n+iL)\exp\left(\frac{-j2\pi kn}{N}\right) \quad (1)$$

where, i indexes the frame number, k represents the frequency bin index corresponding to the frequency $f(k)=kf_s/N$, f_s specifies the sampling frequency, w is a Hamming window function, and L determines the frame shift in samples. Speech spectrum $X(i,k)$ is passed through an auditory filterbank, which resembles the frequency resolution of the human auditory system, after that the number of spectral components reduces and we can get an auditory power spectrum

$$X_{FB}^2(i,j)=\sum_{k=0}^{N-1}\left|h_{FB}(j,k)\times X(i,k)\right|^2 \quad (2)$$

Where, $j=1,2,\ldots,M$, where $M=32$ is the number of auditory filters and $h_{FB}(k,j)$ is a matrix containing the frequency-dependent auditory filter weights. The center frequencies f_c of the auditory filterbank are equally distributed on the equivalent rectangular bandwidth (ERB) scale using a spacing of 1 ERB between 80Hz and 5000Hz. The set of triangular auditory filter weights is computed as

$$h_{FB}(k,j)=\begin{cases}0, & for\ f(k)<f_c(j-1)\\ \dfrac{f(k)-f_c(j-1)}{f_c(j)-f_c(j-1)}, & for\ f_c(j-1)\le f(k)<f_c(j)\\ \dfrac{f(k)-f_c(j+1)}{f_c(j)-f_c(j+1)}, & for\ f_c(j)\le f(k)<f_c(j+1)\\ 0, & for\ f(k)\ge f_c(j+1).\end{cases} \quad (3)$$

After that, we get T-F representation, the auditory power spectrum is loudness compressed by raising it to the power of 0.33 to obtain the spectral features which are used for recognition [6].

2.2. NMF-SNRDA (Nonnegative Matrix Factorization-SNR Discrimination Algorithm)

Many researchers add various constrains in NMF and use it to noise separation and speaker recognition [7-9]. In this paper, firstly use FastICA to separate noise, and then use the result as initial value of NMF，and adds discriminating constrain in NMF to separate pure speech from noisy speech more accurately [10].

Noisy speech spectrum $X(i,k)$ and estimated clean speech spectral $\hat{S}(i,k)$ are obtained through the above process，the estimated noise spectral $\hat{N}(i,k)$ is obtained through spectral subtraction[11-12]. Both $\hat{S}(i,k)$ and $\hat{N}(i,k)$ are transformed to the auditory domain in analogy to（2）:

$$\hat{S}_{FB}^{2}(i,j)=\sum_{k=0}^{N-1}\left|h_{FB}(k,j)\times\hat{S}(i,k)\right|^{2} \quad (4)$$

$$\hat{N}_{FB}^{2}(i,j)=\sum_{k=0}^{N-1}\left|h_{FB}(k,j)\times\hat{N}(i,k)\right|^{2} \quad (5)$$

Classify speech frames by formula (6),

$$m(i,j)=\begin{cases}1, & if\ 10\log_{10}\dfrac{\hat{S}_{FB}^{2}(i,j)}{\hat{N}_{FB}^{2}(i,j)}>LC\\ 0, & otherwise.\end{cases} \quad (6)$$

The value of LC determines the classification results of speech frames [13].

2.3. Differences Detection and Discrimination Algorithm (DDADA)

Short-time energy is often used for evaluating speech frame quality, because it is the most intuitive and easier calculate, but the performance decline in low SNR, in this paper differences detecting and discrimination algorithm is proposed, the algorithm has better robustness in low signal-to-noise.

Every frame of the input signal must get its energy spectrum through the FFT. The description of the algorithm for speech signal based on two assumptions: (1) the speech signal is stable; (2) the spectral energy of each FFT dot obeys Gauss distribution. Therefore, use a Multi- dimension Gauss distribution $S(\mu,\Sigma)$ to describe the spectral character of speech signal. Among them, μ is the mean vector of voice frame energy, Σ is the covariance matrix. In order to reduce the computation cost, assumed Σ to be diagonal matrix, the speech model can be expressed as $S(\mu,\sigma^{2})$. If each frame get the frequency of N point through short time FFT, then

$$\mu = \left(\mu_1, \mu_2, \mu_3, \cdots, \mu_N \right)' \quad (7)$$

$$\sigma^2 = \left(\sigma_1^2, \sigma_2^2, \sigma_3^2, \cdots, \sigma_N^2 \right)' \quad (8)$$

Due to the background noise of speech are complicated, and the characteristics of noise environment have no prior knowledge , at the same time speech recognition must satisfy the real-time requirement, too long time does not permit, to reserve certain frames as pure speech before testing, used to initialize the detection model. After that, according to the detection model calculate the similarity evaluation of each frame of speech. If the spectrum features of input frame is similar to pure speech, the similarity evaluation of the frame is higher, otherwise, is lower. Evaluation of each frame of the input signal can be expressed as:

$$score\left(O_i\right) = S\left(O_i; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left| -\frac{\left(O_i - \mu\right)^2}{\sigma^2} \right| \quad (9)$$

In actual calculation, use the formula (10) to instead of (9).

$$score\left(O_i\right) = \frac{\left(O_i - \mu\right)^2}{\sigma^2} + \ln \sigma^2 = \sum_{n=1}^{N} \left| \frac{\left(O_{i,n} - \mu_n\right)^2}{\sigma_n^2} + \ln \sigma_n^2 \right| \quad (10)$$

Where, $O_i = \left(O_{i,1}, O_{i,2}, O_{i,3}, \cdots, O_{i,J}\right)'$ is the energy spectrum vector of the current frame, the evaluation value is an important feature to differentiate the quality of speech frames.

Several frame reserved is used to initialize the detection model, in order to make detection model better react statistical characteristics of pure speech information, determine the quality of speech frames when detecting the voice signal endpoint. If the analysis result show the current frame is low quality, then the frame is classified to low quality class. If the analysis result shows the current frame is high quality, then the frame is classified to high quality class, and uses the energy spectrum of the current frame to update the detection model. The update process is an iterative process, which makes the detection model be more close to the pure speech model. The update process can be expressed as:

$$\mu_{m+1} = \frac{m\mu_m + S_{m+1}}{m+1} \quad (11)$$

$$\sigma_{m+1}^2 = \frac{\left(m-1\right)\sigma_m^2 + \left(S_{m+1} - \mu_m\right)^2}{m} - \left(\mu_{m+1} - \mu_m\right)^2 \quad (12)$$

μ_{m+1}、σ_{m+1}^2 and μ_m、σ_m^2 respectively are mean vector and variance vector of the speech before updated and after updated; m is the speech frames before update; S_{m+1} is the energy spectral vector of voice frame after updated. Based on evaluation of each frame, eliminate low quality frame, keep high quality frame, to update the clean speech model.

3. GMM-UBM Three-stage Classification Model

GMM (Gaussian Mixture Models) is a multidimensional probability density function, use weighted combination of Gauss distribution probability density function to describe the distribution of vectors in the space of probability density. GMM is used to estimate the D-dimensional feature vector $\vec{x}$ for the task of speaker recognition. Assuming K diagonal Gaussian mixture components, the probability density function of a GMM is given by [14]:

$$p\left(\vec{x} \mid \lambda\right)=\sum_{c=1}^{K} w_c \prod_{m=1}^{D} N\left(x_m, \mu_{c,m}, \sigma_{c,m}^2\right) \quad (13)$$

Where, w_c is the component weight and $N\left(x_m, \mu_{c,m}, \sigma_{c,m}^2\right)$ is a uni-variate Gaussian distribution with mean $\mu_{c,m}$ and variance $\sigma_{c,m}^2$

$$N\left(x_m, \mu_{c,m}, \sigma_{c,m}^2\right)=\frac{1}{\sqrt{2\pi\sigma_{c,m}^2}} \exp\left(-\frac{\left(x_m-\mu_{c,m}\right)^2}{2\sigma_{c,m}^2}\right) \quad (14)$$

The model for each specific speaker can be summarized by the following set of parameters

$$\lambda=\left(w_c, \vec{\mu}_c, \vec{\sigma}_c^2\right) \qquad c=1,\ldots,K. \quad (15)$$

In the two-stage classification model, feature vectors $\vec{x}$ is classified into two sub vectors, which are reliable R and unreliable U. In the process of identification, the two sets are used for speaker recognition respectively. Reliable R is used directly to estimate the similarity score of the speaker λ. We assume that unreliable components are polluted by additive noise, but they do contain information about the maximum energy of the target speech component, so deal with them for recognition.

$$p\left(\vec{x} \mid \lambda\right)=\sum_{c=1}^{K} w_c \prod_{r\in R} N\left(x_r, \mu_{c,r}, \sigma_{c,r}^2\right) \times \prod_{u\in U} \frac{1}{x_{high,u}-x_{low,u}} \int_{x_{low,u}}^{x_{high,u}} N\left(x_u, \mu_{c,u}, \sigma_{c,u}^2\right) dx_u \quad (16)$$

The integral in (16) can be evaluated as the vector difference of error function, and (16) can be rewritten as (17). The bounds are set to $\left[x_{low,u}-x_{high,u}\right]=\left[0, x_u\right]$.

$$p\left(\vec{x} \mid \lambda\right)=\sum_{c=1}^{K} w_c \prod_{r\in R} N\left(x_r, \mu_{c,r}, \sigma_{c,r}^2\right) \times \prod_{u\in U} \frac{1}{x_{high,u}-x_{low,u}} \frac{1}{2}\left[erf\left(\frac{x_{high,u}-\mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right)-erf\left(\frac{x_{low,u}-\mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right)\right] \quad (17)$$

Spectral features are computed using the short-time Fourier transform (STFT) on a frame-by-frame basis. The signal is transformed into overlapping segments and the N-point STFT is computed as

$$X\left(i,k\right)=\sum_{n=0}^{N-1} w\left(n\right) x\left(\left(i-1\right)L+n\right) \exp\left(\frac{-j2\pi kn}{N}\right) \quad (1)$$

where, i indexes the frame number, k represents the frequency bin index corresponding to the frequency $f(k) = kf_s / N$, f_s specifies the sampling frequency, w is a Hamming window function, and L determines the frame shift in samples. Speech spectrum $X(i,k)$ is passed through an auditory filterbank, which resembles the frequency resolution of the human auditory system, after that the number of spectral components reduces and we can get an auditory power spectrum

$$X_{FB}^2(i,j) = \sum_{k=0}^{N-1} \left| h_{FB}(k,j) \times X(i,k) \right|^2 \quad (2)$$

where, $j = 1, 2, \ldots, M$, where $M = 32$ is the number of auditory filters and $h_{FB}(k,j)$ is a matrix containing the frequency-dependent auditory filter weights. The center frequencies f_c of the auditory filterbank are equally distributed on the equivalent rectangular bandwidth (ERB) scale using a spacing of 1 ERB between 80Hz and 5000Hz. The set of triangular auditory filter weights is computed as

$$h_{FB}(k,j) = \begin{cases} 0, & for\ f(k) < f_c(j-1) \\ \dfrac{f(k) - f_c(j-1)}{f_c(j) - f_c(j-1)}, & for\ f_c(j-1) \le f(k) < f_c(j) \\ \dfrac{f(k) - f_c(j+1)}{f_c(j) - f_c(j+1)}, & for\ f_c(j) \le f(k) < f_c(j+1) \\ 0, & for\ f(k) \ge f_c(j+1). \end{cases} \quad (3)$$

Use CNMF to get estimated clean speech spectral $\hat{S}(i,k)$, and then use spectral subtraction [13-14] to get the estimated noise spectral $\hat{N}(i,k)$. Both $\hat{S}(i,k)$ and $\hat{N}(i,k)$ are transformed to the auditory domain in analogy to (2):

$$\hat{S}_{FB}^2(i,j) = \sum_{k=0}^{N-1} \left| h_{FB}(k,j) \times \hat{S}(i,k) \right|^2 \quad (4)$$

$$\hat{N}_{FB}^2(i,j) = \sum_{k=0}^{N-1} \left| h_{FB}(k,j) \times \hat{N}(i,k) \right|^2 \quad (5)$$

Classify speech frames by formula (6),

$$m(i,j) = \begin{cases} 1, & if\ 10\log_{10} \dfrac{\hat{S}_{FB}^2(i,j)}{\hat{N}_{FB}^2(i,j)} > LC \\ 0, & otherwise. \end{cases} \quad (6)$$

The value of LC determines the classification results of speech frames.

3.2. Differences Detection and Discrimination Algorithm (DDADA)

Short-time energy is often used for evaluating speech frame quality, because it is the most intuitive and easier calculate, but the performance decline in low SNR, in this paper differences detecting and discrimination algorithm is proposed, the algorithm has better robustness in low signal-to-noise.

Every frame of the input signal must get its energy spectrum through the FFT. The description of the algorithm for speech signal based on two assumptions: (1) the speech signal is stable; (2) the spectral energy of each FFT dot obeys Gauss distribution. Therefore, use a Multi- dimension Gauss distribution $S(\mu,\Sigma)$ to describe the spectral character of speech signal. Among them, μ is the mean vector of voice frame energy, Σ is the covariance matrix. In order to reduce the computation cost, assumed Σ to be diagonal matrix, the speech model can be expressed as $S(\mu,\sigma^2)$. If each frame get the frequency of N point through short time FFT, then

$$\mu=(\mu_1,\mu_2,\mu_3,\cdots,\mu_N)' \tag{7}$$

$$\sigma^2=(\sigma_1^2,\sigma_2^2,\sigma_3^2,\cdots,\sigma_N^2)' \tag{8}$$

The background noise of speech are complicated and the characteristics of noise environment have no prior knowledge , at the same time speech recognition must satisfy the real-time requirement, too long time does not be permitted, so reserve first certain frames as pure speech before testing, being used to initialize the detection model. After that, according to the detection model calculate the similarity evaluation of each frame of speech. If the spectrum features of input frame is similar to pure speech, the similarity evaluation of the frame is higher, otherwise, is lower. Evaluation of each frame of the input signal can be expressed as:

$$score(O_i)=S(O_i;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left|-\frac{(O_i-\mu)^2}{\sigma^2}\right| \tag{9}$$

In actual calculation, use the formula (4) to instead of (3).

$$score(O_i)=\frac{(O_i-\mu)^2}{\sigma^2}+\ln\sigma^2=\sum_{n=1}^{N}\left|\frac{(O_{i,n}-\mu_n)^2}{\sigma_n^2}+\ln\sigma_n^2\right| \tag{10}$$

$O_i=(O_{i,1},O_{i,2},O_{i,3},\cdots,O_{i,J})'$ is the energy spectrum vector of the current frame, the evaluation value is an important feature to differentiate the quality of speech frames.

Several frame reserved is used to initialize the detection model, in order to make detection model better react statistical characteristics of pure speech information, determine the quality of speech frames in detecting the voice signal endpoint. If the analysis result show the current frame is low quality, then the frame is discarded, if the analysis result shows the current frame is high quality, and then uses the energy spectrum of the current frame to update the detection model. The update process is an iterative process, which makes the detection model be more close to the pure speech model. The update process can be expressed as:

$$\mu_{m+1} = \frac{m\mu_m + S_{m+1}}{m+1} \quad (11)$$

$$\sigma_{m+1}^2 = \frac{(m-1)\sigma_m^2 + (S_{m+1} - \mu_m)^2}{m} - (\mu_{m+1} - \mu_m)^2 \quad (12)$$

μ_{m+1}、 σ_{m+1}^2 and μ_m、 σ_m^2 respectively are mean vector and variance vector of the speech before updated and after updated; m is the speech frames before update; S_{m+1} is the energy spectral vector of voice frame after updated. Based on evaluation of each frame, eliminate low quality frame, keep high quality frame, to update the clean speech model.

4 .GMM-UBM Three-stage Classification Model

GMMs (Gaussian Mixture Models) is a multidimensional probability density function, use weighted combination of Gauss distribution probability density function to describe the distribution of vectors in the space of probability density. GMMs are used to estimate the D-dimensional feature vector $\vec{x}$ for the task of speaker recognition. Assuming K diagonal Gaussian mixture components, the probability density function of a GMMs is given by formula (13):

$$p(\vec{x} \mid \lambda) = \sum_{c=1}^{K} w_c \prod_{m=1}^{D} N(x_m, \mu_{c,m}, \sigma_{c,m}^2) \quad (13)$$

Where, w_c is the component weight and $N(x_m, \mu_{c,m}, \sigma_{c,m}^2)$ is a uni-variate Gaussian distribution with mean $\mu_{c,m}$ and variance $\sigma_{c,m}^2$

$$N(x_m, \mu_{c,m}, \sigma_{c,m}^2) = \frac{1}{\sqrt{2\pi\sigma_{c,m}^2}} \exp\left(-\frac{(x_m - \mu_{c,m})^2}{2\sigma_{c,m}^2} \right) \quad (14)$$

The model for each specific speaker can be summarized by the following set of parameters.

$$\lambda = (w_c, \vec{\mu}_c, \vec{\sigma}_c^2) \qquad c = 1, \ldots, K. \quad (15)$$

In the two-stage classification model, feature vector $\vec{x}$ is classified into two sub vectors, which are reliable R and unreliable U. In the process of identification, the two sets are used for speaker recognition respectively. Reliable R is used directly to estimate the similarity score of the speaker λ.

We assume that unreliable components are polluted by additive noise, but they do contain information about the maximum energy of the target speech component, so deal with them for recognition.

$$p(\vec{x} \mid \lambda) = \sum_{c=1}^{K} w_c \prod_{r \in R} N(x_r, \mu_{c,r}, \sigma_{c,r}^2) \times \prod_{u \in U} \frac{1}{x_{high,u} - x_{low,u}} \int_{x_{low,u}}^{x_{high,u}} N(x_u, \mu_{c,u}, \sigma_{c,u}^2) dx_u \quad (16)$$

The integral in （16） can be evaluated as the vector difference of error function, and （16） can be rewritten as （17）. The bounds are set to $\left[x_{low,u}-x_{high,u}\right]=\left[0,x_u\right]$.

$$p\left(\vec{x}\mid\lambda\right)=\sum_{c=1}^{K}w_c\prod_{r\in R}N\left(x_r,\mu_{c,r},\sigma_{c,r}^2\right)\times\prod_{u\in U}\frac{1}{x_{high,u}-x_{low,u}}\frac{1}{2}\left[erf\left(\frac{x_{high,u}-\mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right)-erf\left(\frac{x_{low,u}-\mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right)\right] \tag{17}$$

In the three-stage classification model, feature vectors $\vec{x}$ is classified into three sub vectors, which are high quality R, medium quality M and low quality U. In the process of identification, the three sets are used for speaker recognition respectively. High quality R is used directly to estimate the similarity score of the speaker λ. We assume that medium quality M and low quality U components are polluted by different degree additive noise, but they do contain information about the maximum energy of the target speech component, so deal with them for recognition respectively.

$$\begin{aligned}p\left(\vec{x}\mid\lambda\right)=&\sum_{c=1}^{K}w_c\prod_{h\in H}N\left(x_h,\mu_{c,h},\sigma_{c,h}^2\right)\times\prod_{l\in L}\frac{1}{x_{high,l}-x_{low,l}}\int_{x_{low,l}}^{x_{high,l}}N\left(x_l,\mu_{c,l},\sigma_{c,l}^2\right)dx_l\\&\times\prod_{m\in M}\frac{1}{2}\left[N\left(x_m,\mu_{c,m},\sigma_{c,m}^2\right)+\frac{1}{x_{high,m}-x_{low,m}}\int_{x_{low,m}}^{x_{high,m}}N\left(x_m,\mu_{c,m},\sigma_{c,m}^2\right)\right]\end{aligned} \tag{18}$$

In practical computation,（18）can be rewritten as（19）, the bounds are set to $\left[x_{low,l}-x_{high,l}\right]=\left[0,x_l\right]$, $\left[x_{low,m}-x_{high,m}\right]=\left[0,x_m\right]$.

$$\begin{aligned}p\left(\vec{x}\mid\lambda\right)=&\sum_{c=1}^{K}w_c\prod_{h\in H}N\left(x_h,\mu_{c,h},\sigma_{c,h}^2\right)\times\prod_{l\in L}\frac{1}{x_{high,l}-x_{low,l}}\frac{1}{2}\left[erf\left(\frac{x_{high,l}-\mu_{c,l}}{\sqrt{2\sigma_{c,l}^2}}\right)-erf\left(\frac{x_{low,l}-\mu_{c,l}}{\sqrt{2\sigma_{c,l}^2}}\right)\right]\\&\times\prod_{m\in M}\frac{1}{2}\left\{N\left(x_m,\mu_{c,m},\sigma_{c,m}^2\right)+\frac{1}{x_{high,m}-x_{low,m}}\frac{1}{2}\left[erf\left(\frac{x_{high,m}-\mu_{c,m}}{\sqrt{2\sigma_{c,m}^2}}\right)-erf\left(\frac{x_{low,m}-\mu_{c,m}}{\sqrt{2\sigma_{c,m}^2}}\right)\right]\right\}\end{aligned} \tag{19}$$

5. Experiments and Results

5.1. Speech Database and Noise

The speech database is the TIMIT speech database, the sampling rate is 16 KHz, mono recording, 16Bit quantification, including 630 speakers, contains two subdirectories; Train directory and Test directory. Each directory contains 8 folders from Dr1 to Dr8, the eight folders represent eight different dialects of English, each speaker read 10 statements, and the length of each sentence is about 3 seconds.

Experimental samples were obtained from TIMIT speech database added noise, we used 300 speakers of them; the training corpus taken from the first sentence of each speaker, the length is about 3 seconds; the test data taken from each speaker's tenth words, it is about 2 seconds. The complex noise under battlefield environment is added to each speech according to different SNR.

5.2. The Relationship between Classification Methods and Recognition Rate

In this experiment, we use four methods to classify speech frames.

Method 1: Endpoint detection and discrimination algorithm (EDADA), get noise fragment by the endpoint detection, and then get clean speech spectral by spectral subtract, classify the speech frame by formula (6).

Method 2: Improvement SNR discrimination algorithm (ISNRDA), get clean speech spectral by FastICA separation algorithm and get noise spectral by spectral subtract, then classify the speech frame by formula (6).

Method 3: Use DDADA proposed in this paper to classify the speech frame.

Method 4: Use NMF-SNRDA proposed in this paper to classify the speech frame.

Table 1. Combines the Four Classification Methods with GMM-UBM Two-stage Classification Model

SNR（dB）				
Method	-5	0	5	10
EDADA	27.333%	33.667%	38.667%	47.000%
ISNRDA	34.333%	43.667%	49.333%	53.667%
DDADA	40.667%	47.333%	52.000%	55.333%
NMF-SNRDA	42.000%	51.667%	55.333%	57.667%

Table 1 shows that NMF-SNRDA discrimination algorithm can make the best recognition performance.

5.3. Research on the Dual Information Quality Discrimination Algorithm

Table 2. Relationship of the Dual Information Quality Discrimination Algorithm and Noisy Short Utterance Recognition Rate (Combine with GMM-UBM Two-stage Classification Model)

SNR（dB）				
Method	-5	0	5	10
EDADA +DDADA	37.667%	46.000%	51.333%	55.000%
ISNRDA +DDADA	44.333%	52.667%	56.667%	57.333%
NMF-SNRDA+ ISNRDA	46.667%	54.333%	57.000%	58.667%
NMF-SNRDA+DDADA	48.667%	56.667%	58.333%	60.000%

Table 3. Relationship of the Dual Information Quality Discrimination Algorithm and Noisy Short Utterance Recognition Rate (Combine with GMM-UBM Three-stage Classification Model)

SNR（dB）				
Method	-5	0	5	10
EDADA +DDADA	41.667%	49.000%	53.667%	57.333%
ISNRDA +DDADA	48.000%	55.333%	57.667%	58.667%
NMF-SNRDA+ ISNRDA	50.000%	57.333%	59.333%	60.000%
NMF-SNRDA+DDADA	51.333%	58.667%	60.667%	61.333%

Table 2 and table 3 show that combining all kinds of methods with GMM-UBM three-stage classification model can improve the speaker recognition rate in different SNR, comparing with combining all kinds of methods with GMM-UBM two-stage classification model. NMF-SNRDA+DDADA can get the highest recognition rate in all kinds of methods.

6. Conclusion

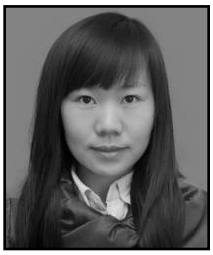
The noisy short utterance is polluted by noise and its corpus is less, so the recognition rate of the noisy short utterance is bad. In this paper, we proposed the corresponding compensation methods. We proposed the algorithm of NMF-SNRDA+DDADA, it can classify speech frames more accurately and it can reduce the influence of noise on speaker recognition rate the noisy short utterance. We proposed GMM-UBM three-stage classification model, it can help make full use of limited corpus. Finally, we combine the algorithm of NMF-SNRDA+DDADA with GMM-UBM three-stage classification model. The experiments confirmed that the above algorithms can make full use of limited corpus and classify speech frames more precisely, and the above algorithms can improve speaker recognition rate of noisy short speech.

References

[1] J. Ye and Z. Tang, “Research on the speaker identification based on short utterance”, Journal of Electronics, vol. 4, no. 39, (**2011**).
[2] F. Nakhat and T. F. Zheng, “Short utterance speaker recognition. International Conference on Systems and Informatics (ICSAI2012)”, (2012) May19-20, Beijing, China.
[3] M. Tobias, S. van de Par and A. Kohlrausch, “Noise-robust speaker recognition combining missing data techniques and universal background modeling [J]”, IEEE Transactions on Audio, Speech and Language Processing, vol. 1, no. 20, (**2012**).
[4] N. Dehak, “Front-end factor analysis for speaker verification [J]”, IEEE Transactions on Audio, Speech and Language Processing, vol. 4, no. 19, (**2011**).
[5] A. Kanagasundaram and R. Vogt, “i-vector based speaker recognition on short utterance”, Conference of the International Speech Communication Association (InterSpeech), (**2011**) August 27-31, Florence, Italy.
[6] T. May, S. v. de Par and A. Kohlrausch, “Noise-robust speaker recognition combining missing data techniques and universal background modeling [J]”, IEEE Transactions on Audio,Speech and Language Processing, vol. 1, no. 20, (**2012**).
[7] Y. Tian, Y. Li, H. Lin and H. Ma, “A sparse NMF-SU for seismic random noise attenuation”, Geoscience and Remote Sensing Letters, IEEE, vol. 3, no. 10, (**2013**).

[8] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization", 10th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA (2012) March 12-15, Tel Aviv, Israel.
[9] C. Joder and B. Schuller, "Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition", ITG Conference on Speech Communication, IEEE, (2012) September 26-28, Braunschweig Germany.
[10] H. F. Liu, Z. H. Wu and D. Cai, "Constrained Nonnegative matrix factorization for image representation [J]", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 7, no. 34, (**2012**).
[11] G. Paurav and G. Anil, "Developments in spectral subtraction for speech enhancement [J]", International Journal of Engineering Research and Application, vol. 1, no. 2, (**2012**).
[12] Ch. V. Rama, M. B. Rama and K. Srinivase, "Noise reduction using mel-scale spectral subtraction with perceptually defined subtraction parameters-a new scheme [J]", Signal and Image Processing, vo. 1, no. 2, (**2011**).
[13] M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data", Speech Communication, vol. 3, no. 34, (**2001**).

Authors



Ying Chen, she received her Master's degree in Northeast Dianli University in Jilin, China and is currently a Ph.D. student in Nanjing University of Science and Technology. Her research interest is mainly in the area of noisy short utterance signal processing and she has published several research papers in scholarly journals and international conferences in the above research areas.

Zhenmin Tang, he received his M.S. degree in East China Institute of Technology and Ph.D. degree in Nanjing University of Science and Technology. He is currently a professor in the School of Computer Science & Engineering, Nanjing University of Science and Technology, China. He's research interest is mainly in the area of speech recognition, image processing and intelligent robot. He has published several hundreds of research papers in scholarly journals and international conferences in the above research areas.

