International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.4 (2014), pp.145-154
http://dx.doi.org/10.14257/ijsip.2014.7.4.14

Research on Noise Processing and Speech Frame Classification Model of Noisy Short Utterance Signal Processing

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 noise processing is the key of improving recognition rate for the noisy utterance. While for the short utterance, its corpus is less and small amount of speech data is available for testing and training, so making full use of its limit corpus is the key of improving recognition rate of the short utterance. For the noisy short utterance, the noise processing and making full use of the limit corpus are vital. We proposed noise separation algorithm based on constrained Non-negative matrix factorization (CNMF) to make the noise processing. As making full use of the limit corpus, we proposed the improved SNR discrimination algorithm (ISNRDA) and the differences detection and discrimination algorithm (DDADA), we use the two classification algorithm to estimate the quality of the speech frame, and classify the speech frame. Besides, we combine the above classification result with the GMM-UBM three-stage classification model proposed in this paper, so that we can make full use of the limit corpus of the noisy short utterance. Experiments show that the above algorithms can improve speaker recognition performance of noisy short utterance.

Key Words: *noisy short utterance, CNMF, ISNRDA, DDADA, three-stage classification model*

1. Introduction

The research on the signal processing of noisy short utterance is a very important issue in current era of information technology. Research on the signal processing of noisy short utterance is complicated and real, and it originated in the signal processing of speaker recognition. Now the field of speaker recognition has made a lot of research achievements [1-3]. Joint factor analysis technology recently proposed by Kenny (JFA) 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]. Short Utterance Speaker Recognition (SUSR) is an important area of speaker recognition when only small amount of speech data is available for testing and training [6]. While in the real environment, short utterance is usually affected by noise, and noise can make the recognition rate reduce significantly. So, noise processing is very important for noisy speech signal [7] and separation of clean speech and background noise and accurate judgment of the reliability of speech frames are helpful to improve the speaker recognition rate of noisy short utterance [8-10]. In many studies of noisy short utterance, researchers use NMF algorithm to improve speaker recognition rate [11-12].

We propose noise separation algorithm based on constrained Non-negative matrix

ISSN: 2005-4254 IJSIP

factorization (CNMF) to make the noise processing. As making full use of the limit corpus, we proposed the improved SNR discrimination algorithm (ISNRDA) and the differences detection and discrimination algorithm (DDADA), we use the two classification algorithm to estimate the quality of the speech frame, and classify the speech frame. Besides, we combine the above classification result and the GMM-UBM three-stage classification model proposed by us.

2. Noise Processing Algorithm

2.1. Fast ICA Algorithm

Blind source separation algorithm is widely used in the separation of mixed speech signals, obtained the good separation effect, an important part of blind source separation is Fast ICA algorithm based on maximum entropy theory, it is a kind of fast separation algorithm based on fixed-point iteration, the algorithm not only has good stability, and fast convergence.

2.2. Constrained Non-negative Matrix Factorization (CNMF)

Recently, many researchers add constrains in NMF and use it to separate noise and make speaker recognition. We use Fast ICA to initialize NMF, and add discriminating constrain to NMF.

Given a non-negative matrix$W = [x_1, x_2, \cdots, x_n] \epsilon R^{m \times n}$, W can be expressed as the product of two non-negative matrices, W=UV.

Assuming that the first l samples are labeled in the data set $\{x_i\}_{i=1}^{n}$, the remaining $n-l$ samples are not labeled, the data set consists of c samples classes. If the sample x_i belongs to the jth class sample, then$c_{ij} = 1$, otherwise$c_{ij} = 0$, so we define a constrained classification matrix $A = \begin{pmatrix} C_{l \times c} & 0 \\ 0 & I_{n-l} \end{pmatrix}$, I_{n-l} is a $(n-l)(n-l)$ -dimensional unit matrix. Add classification information to non-negative matrix Z, we can get $V = AZ$, where $Z \epsilon R^{(c+n-l) \times k}$, $k \ll m$ and $k \ll n$. If x_i and x_j belong to the same sample class, then$v_i = v_j$, and $X \approx U(AZ)^T$

3. Speech Frame Classification Algorithm

Firstly, we use noise processing algorithm to process noisy utterance, after that, we use speech frame classification algorithm to classify the speech frames.

3.1. Improved SNR Discrimination Algorithm (ISNRDA)

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

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

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 filter bank, 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, $j=1,2,\cdots,M$, 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 filter bank 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

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

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

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:

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

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

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

$O_i = (O_{i,1}, O_{i,2}, O_{i,3}, \cdots, O_{i,J})'$ 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.

$$\lambda = (w_{c}, \vec{\mu}_{c}, \vec{\sigma}_{c}^{2}) \qquad c = 1, \cdots, K. \quad (15)$$

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 $[x_{low,u} - x_{high,u}] = [0, x_{u}]$.

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

In the three-stage classification model, feature vector $\vec{x}$ 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.

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

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

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

Speaker-independent UBM is trained on the pooled speech material of many speakers using k-means clustering and EM algorithm. A speaker-dependent model is derived by adapting the well-training the UBM parameters to the speech material of the corresponding speaker using maximum a posteriori (MAP) estimation.

During the adaptation process, only those Gaussian components of the UBM are adapted, which show sufficient probabilistic alignment with the speaker-dependent speech material. In this way, the parameters of Gaussian components which are potentially under-represented are not updated to the new data, making the model adaptation robust even to a small amount of training data. The MAP adaptation was shown to outperform the estimation of GMM parameters using the maximum-likelihood approach.

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. The complex noise under battlefield environment is added to each speech according to different SNR.

Experimental samples were obtained from TIMIT speech database added noise, we used 230 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.

5.2. Research on Noise Separation Algorithm

Experimental methods: use the following three methods to initialize noisy short speech, respectively.

Method 1: use FastICA algorithm to initialize noisy short speech.

Method 2: use random initialization NMF algorithm to initialize noisy short speech.

Method 3: use CNMF algorithm to initialize noisy short speech.

After initialize noisy short utterance, we extract MFCC features and combine MFCC feature with the GMM-UBM model. Table 1 shows the relationship between three methods of noisy separation and speaker recognition rate of noisy short speech.

Table 1. Research on Noise Separation Algorithm

	SNR（dB）			
methods	0	5	10	15
Method 1	33.478%	41.304%	46.521%	48.260%
Method 2	30.869%	38.260%	44.347%	46.086%
Method 3	36.521%	43.913%	49.130%	50.869%

Table 1 shows that noise separation algorithm based on CNMF proposed in this paper can make the highest recognition rate in the three methods.

5.3. Research on the Quality Discrimination Algorithm of the Speech Frame

Experimental methods: use the following three methods to classify the speech frame, respectively.

Method 4: use FastICA algorithm to initialize noisy short speech, and then use ISNRDA to classify the speech frame.

Method 5: use random initialization NMF algorithm to initialize noisy short speech, and then use ISNRDA to classify the speech frame.

Method 6: use CNMF algorithm to initialize noisy short speech, and then use ISNRDA to classify the speech frame.

After that, we extract MFCC feature and combine MFCC feature with the GMM-UBM two-stage classification model.

Table 2. Research on the Speech Frame Classification Algorithm

	SNR（dB）			
methods	0	5	10	15
Method 4	40.434%	47.391%	52.609%	54.347%

Method 5	37.391%	44.782%	50.000%	52.174%
Method 6	43.391%	49.608%	54.956%	56.260%

Table 2 shows the relationship between the above three methods and speaker recognition rate of noisy short speech.

Experimental methods: use the following three methods to classify the speech frame, respectively.

Method 7: use FastICA algorithm to initialize noisy short speech, and then use DDADA to classify the speech frame.

Method 8: use random initialization NMF algorithm to initialize noisy short speech, and then use DDADA to classify the speech frame.

Method 9: use CNMF algorithm to initialize noisy short speech, and then use DDADA to classify the speech frame.

After that, we extract MFCC feature and combine MFCC feature with the GMM-UBM two-stage classification model.

Table 3 shows the relationship between the above three methods and speaker recognition rate of noisy short speech.

Table 3. Research on the Speech Frame Classification Algorithm

SNR（dB）				
methods	0	5	10	15
Method 7	44.348%	49.130%	53.913%	56.087%
Method 8	42.609%	47.391%	52.174%	54.783%
Method 9	47.826%	53.043%	57.391%	58.260%

Table 2 and Table 3 show that noise separation algorithm based on CNMF proposed in this paper can make the highest recognition rate in the three noise separation algorithms, and the two discrimination algorithms--ISNRDA and DDADA can improve speaker recognition rate of noisy short utterance in different degree.

5.4. Research on the Speech Frame Classification Model

Experimental methods:

Method 10: use CNMF algorithm to initialize noisy short speech, and then meanwhile use ISNRDA and DDADA to classify the speech frame, If the two discrimination algorithm determine the same speech frame is low quality, the speech frames is classified to low quality class, if not the speech frames is classified to high quality class.

Method 11: use CNMF algorithm to initialize noisy short speech, and then meanwhile use ISNRDA and DDADA to classify the speech frame, If the two discrimination algorithm determine the same speech frame is high quality, the speech frames is classified to high quality class, if not the speech frames is classified to low quality class.

Method 12: use CNMF algorithm to initialize noisy short speech, and then meanwhile use ISNRDA and DDADA to classify the speech frame, If the two discrimination algorithm determine the same speech frame is high quality, the speech frames is classified to high quality class; If the two discrimination algorithm determine the same speech frame is low quality, the speech frames is classified to low quality class; if not the speech frames is classified to medium quality class.

After that, we extract MFCC feature and combine MFCC feature with the GMM-UBM three-stage classification model.

Table 4. Research on the Speech Frame Classification Model

	SNR（dB）			
methods	0	5	10	15
Method 10	51.304%	56.087%	59.565%	61.304%
Method 11	54.348%	58.261%	60.870%	62.609%
Method 12	58.261%	62.174%	63.913%	65.217%

Table 4 shows that three-stage classification model can improve speaker recognition rate of noisy short utterance, comparing with two-stage classification model.

6. Conclusion

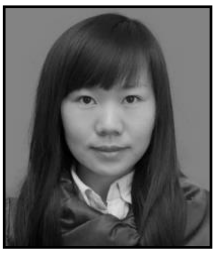
The mainly factors impacting speaker recognition performance of noisy short utterance are noise and inadequate corpus. In this paper, we research on the two mainly factors impacting speaker recognition performance of noisy short utterance, and propose corresponding compensation algorithms. We use CNMF algorithm to initialize noisy short speech, and then use ISNRDA and DDADA to classify the speech frame, If the two discrimination algorithm determine the same speech frame is high quality, the speech frames is classified to high quality class; If the two discrimination algorithm determine the same speech frame is low quality, the speech frames is classified to low quality class; if not, the speech frames is classified to medium quality class. After that, we extract MFCC feature and combine MFCC feature with the GMM-UBM three-stage classification model proposed in this paper. The experiments confirmed that the above algorithms can improve speaker recognition rate of noisy short speech.

References

[1] S. Farah and A. Shamim, "Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization", 3rd International Conference on Computer, Control & Communication (IC4), **(2013)** September 25-26, Karachi, Pakistan.
[2] Y. Tsao, S. Matsuda, C. Hori, H. Kashioka and C.-H. Lee, "MAP-based Online Estimation Approach to Ensemble Speaker and Speaking Environment Modeling", IEEE Transactions on Audio, Speech, and Language Processing, vol. 2, no. 22, **(2014)**.
[3] S. O. Sadjadi and J. H. L. Hansen, "Robust front-end processes for speaker identification over extremely degraded communication channels", ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, **(2013)** May 26-31, Vancouver, BC, Canada.
[4] N. Dehak, *et al.*, "Front-end factor analysis for speaker verification", IEEE Transactions on Audio, Speech and Language Processing, vol. 4, no. 19, **(2011)**.
[5] A. Kanagasundaram, R. Vogt, *et al.*, "I-vector based speaker recognition on short utterance", Conference of the International Speech Communication Association (InterSpeech), **(2011)** August 27-31, Florence, Italy.
[6] N. F. T. F. Zheng, "Short utterance speaker recognition. International Conference on Systems and Informatics (ICSAI2012)", **(2012)** May19-20, Beijing, China.
[7] J. H. L. Hansen, "Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition", IEEE Transactions on Audio, Speech and Language Processing, vol. 2, no. 17, **(2009).**
[8] S. Harding, J. Barker and G. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction", IEEE Transactions on Audio, Speech and Language Processing, vol. 1, no. 14, **(2006).**
[9] T. Hasan and J. H. L. Hansen, "Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker

Verification in Noise", IEEE Transactions on Audio, Speech and Language Processing, vol. 2, no. 22, **(2014)**.
[10] T. May，S. Van de Par and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling", IEEE Transactions on Audio, Speech and Language Processing, vol. 1, no. 20, **(2012)**.
[11] 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.
[12] C. Joder, 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.
[13] G. Paurav and G. Anil, "Developments in spectral subtraction for speech enhancement", International Journal of Engineering Research and Application, vol. 1, no. 2, **(2012)**.
[14] R. C. V. Rama, M. M. B. Rama and R. K. Srinivasa, "Noise reduction using mel-scale spectral subtraction with perceptually defined subtraction parameters-a new scheme", Signal and Image Processing, vol. 1, no. 2, **(2011).**

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.

