Improved MFCC Feature Extraction Combining Symmetric ICA Algorithm for Robust Speech Recognition

Huan Zhao, Kai Zhao, He Liu
School of Information Science and Engineering, Hunan University, Changsha, China
Email: hzhao@hnu.edu.cn, zhaokai@hnu.edu.cn, liuhe@hnu.edu.cn

Fei Yu
Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou, China
Email: hunanyufei@126.com

***Abstract*—Independent component analysis (ICA), instead of the traditional discrete cosine transform (DCT), is often used to project log Mel spectrum in robust speech feature extraction. The paper proposed using symmetric orthogonalization in ICA for projecting log Mel spectrum into a new feature space as a substitute in extracting speech features to solve the problem of cumulative error and unequal weights that deflation orthogonalization brings, so as to improve the robustness of speech recognition systems, and increase the efficiency of estimation at the same time. Furthermore, the paper studied the nonlinearities of the objective function in ICA and their coefficients, tested them in all kinds of environments, finding that they influenced the recognition rate greatly in speech recognition systems, and applied a new coefficient in the proposed method. Experiments based on HMM and Aurora-2 speech corpus suggested that the new method was superior to deflation-based ICA and MFCC.**

***Index Terms*—independent component analysis, speech feature extraction, speech recognition**

I. Introduction

Speech feature extraction has been a key focus in robust speech recognition research[1]. Selecting appropriate features guarantees the good performance of a speech recognition system. Among a large amount of methods for speech feature extraction, the ones based on spectrum are widely used, especially Mel frequency cepstral coefficients (MFCC). Although many new methods for feature extraction are proposed constantly, such as non-stationary feature extraction[2], Gabor analysis and tensor factorization based feature extraction[3], etc, MFCC is still the most important method for speech feature extraction in state-of-the-art automatic speech recognition systems.

Because the feature space by DCT is not dependent on real speech data directly, MFCC performs poor in noisy environment. Data-driven feature space transformations are highly adaptable to real speech data, and will achieve better results than DCT in a practical environment. Principle component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA) are frequently-used data-driven linear transformations. These transformations replace DCT in MFCC procedure to transform the feature space of logarithmic spectrum for new speech features. On the basis of the principle of minimum reconstruction error, PCA projects spectral coefficients onto the direction of maximum variance. ICA performs feature transformation based on the hypothesis of statistical independence of independent components, expecting to find the original structure of speech features.

Independent component analysis has become an important method in statistics, and makes significant progress especially in the field of blind source separation[4]. Recently ICA draws more and more attention in speech feature extraction. FastICA method is widely used because of its high efficiency[5], mainly in speech feature extraction when used in speech recognition. When estimating many independent components, there are two decorrelation modes in FastICA: deflation (serial) and symmetric (parallel) orthogonalization method[6]. The paper discussed two different methods in speech feature extraction, and talked about the nonlinearities of objective function in FastICA and their coefficients.

Feature transformation is a common method in speech feature extraction, projecting the feature space in order to achieve decorrelation[7], dimensionality reduction and noise reduction. There are two main categories[8]: linear feature transformations, such as DCT, PCA, LDA, and ICA, etc.; nonlinear feature transformations, such as nonlinear principal component analysis (NPCA), nonlinear discriminant analysis (NLDA), nonlinear

National Science Foundation of China (Grant NO. 61173106), the Key Program of Hunan Provincial Natural Science Foundation of China (Grant No.10JJ2046), the Planned Science and Technology Key Project of Hunan Province, China (Grant No.2010GK2002).

Corresponding author: Huan Zhao(email: hzhao@hnu.edu.cn)

doi:10.4304/jmm.7.1.74-81

independent component analysis (NICA) and so on. Reference [8] applied PCA, LDA, ICA and nonlinear LDA in a phone recognition task using TIMIT, and compared the results of the different speech features. Reference [9] extracted the correlation information of subspace of phones using PCA in order to extracting speech features. Reference [10] transformed some different speech features using LDA, and reduced the recognition error rate efficiently. Taking the computational complexity and accuracy into account, linear feature transformations methods are commonly used, and applied after getting the Log Mel spectrum. DCT is a non-data related transformation, so it can't adapt to the characteristics of the actual data, and achieves only partial decorrelation[11]. LDA determines complexly and is sensitive to the mismatch of SNR of training and testing set. On the basis of the principle of minimum reconstruction error, PCA projects spectral coefficients onto the direction of maximum variance. ICA regards the inputted multidimensional data as a linear combination of independent components and reestimates the original independent components according to some objective, in order to obtain the physical structure and formation of these components[12]. After pre-emphasis, frame windowing, FFT, Mel filtering and logarithms, feature coefficients are gotten. MFCC, PCA features, and ICA features are gotten when applying DCT, PCA and ICA respectively to feature coefficients. Based on deflation and symmetric decorrelation categories, ICA features can be classified into deflation ICA features (ICA_DEFL) and symmetrical ICA features (ICA_SYMM). In the experiments the paper compared the influence of four different features on robustness and accuracy of automatic speech recognition systems. The following of the paper first introduced the ICA principle and described the feature extraction method using ICA. At the same time the paper researched the influence of nonlinearities of objective function and their coefficients on automatic speech recognition systems, and then tested them to verify the performance. Finally, the paper discussed and summarized the experimental results.

II. Feature Extraction Based on Symmetric ICA

A. The Principle of ICA

Independent component analysis (ICA) is a method which finds internal factors or components from multivariate statistical data[12], looking for both statistically independent and non-Gaussian components. ICA is used in blind source separation at the earliest, but recently also applied to feature extraction gradually. In reference [13] the author used ICA to replace the Fourier transform. In reference [11] ICA was applied to log Mel spectrum. Assuming observed random variables $x_1, x_2, ..., x_n$, each of which is a linear combination of another n random variables $s_1, s_2, ..., s_n$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, i = 1, \cdots, n \qquad (1)$$

In (1), $a_{ij}, i, j = 1, \cdots, n$ are real coefficients, assuming all si are statistically independent. Only random variables xi can be observed, aij and si must be estimated just by xi. Eq. (1) can be show using matrix as $x = A * s$. Random vector x represents mixed vector, s represents independent components, and A represents a matrix which is composed of a_{ij}. To obtain independent components, a demixed matrix *W* should be computed:

$$u = W * x \qquad (2)$$

Where *W* is the inverse matrix of matrix A, and u is an estimate of s.

B. Feature Extraction Based on Symmetric ICA

According to different principles, there are various methods to estimate *W* in ICA, such as maximizing the nongaussianity method, maximum likelihood estimation method and minimizing the mutual information method, etc. An important method of maximizing the nongaussianity methods is FastICA. When estimating multiple independent components using FastICA, they can be estimated one by one using deflation orthogonalization algorithm one by one. Each time one vector w_i is initialized, updated, orthogonalized, and normalized until it converges. Independent components also can be estimated using symmetric orthogonalization method. Every w_i is iterated firstly, and then all w_i are orthogonalized using a special way.

Deflation (serial) orthogonalization method and symmetric (parallel) orthogonalization method computes *W* respectively as Fig. 1:

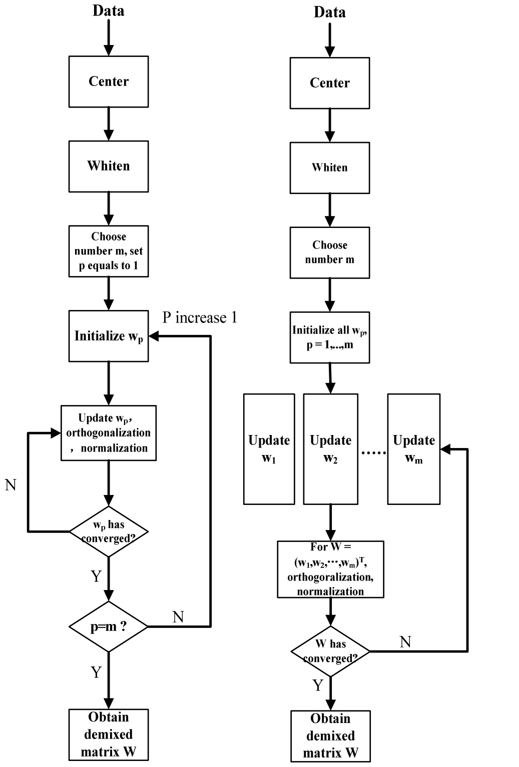


Figure 1. The deflation orthogonalization of ICA and symmetric orthogonalization of ICA.

The difference between the two methods lies in calculating the demixed matrix *W* in different ways. The former calculates each component of W one by one, updates and orthogonalizes them using (3) and (7) respectively until they converge while the other calculates them in parallel, updates and orthogonalizes them using (3) and (8) respectively until they converge.

$$w_p = E\{zg(w_p{}^T z)\} - E\{g'(w_p{}^T z)\}w_p \qquad (3)$$

Where g can be (4), (5) or (6)

$$g_1(y) = \tanh(a1 * y) \qquad (4)$$

$$g_2(y) = y * \exp(-a2 * u^2 / 2) \qquad (5)$$

$$g_3(y) = y^3 \qquad (6)$$

$$w_p \leftarrow w_p - \sum_{j=1}^{p-1} (w_p{}^T w_j) w_j \qquad (7)$$

$$W = (WW^T)^{-1/2} W,$$

where $W = (w_1, w_2, \cdots, w_p)^T$ （8）

There are some deficiencies in deflation orthogonalization method: the error of firstly estimated components will be accumulated so as to influence the estimate of following components which is brought by the orthogonalization. Independent components can be calculated in parallel using symmetric orthogonalization method to solve the above problem, meanwhile the time of calculating $\boldsymbol{W}$ can be shortened sharply. In following experiments, we extracted speech features using the two methods, compared the influence on automatic speech recognition systems which they brought, and discussed the advantages and disadvantages of them.

C. Nonlinearities and their coefficients

The statistical properties of ICA (such as consistence, asymptotic variance, robustness) depend on the selection of objective functions. In objective functions, the non-quadratic functions G are very important, and provide high-level information in the form of expectation $E\{b_i^T x\}$. In actual algorithms, this is equivalent to choosing the derivative of G, nonlinearities g. In the following of the paper, G and g were both referred as nonlinearities. Reference [6] proved that the optimal non-quadratic functions are in the following form:

$$G_{opt}(y) = |y|^a, a < 2 \qquad (9)$$

However, the problem of the above functions is that they are not derivable in origin when a<=1 and this leads to numerical optimization problem. Reference [6] indicated that : (1) A good general-purpose function is $G(y) = \log(\cosh(a1 * y))$, where 1<=a1<=2 and a1 is a real number; (2) When independent components are super-Gaussian or robustness is very important, $G(y) = -\exp(-y^2/2)$ may work well; (3) Only when independent components are sub-Gaussian and there are no outliers, kurtosis is proper. Three functions are as follows:

$$G_1(y) = \frac{1}{a1} \log(\cosh(a1 * y)) \qquad (10)$$

$$G_2(y) = -\exp(-a2 * y^2 / 2) \qquad (11)$$

$$G_3(y) = \frac{1}{4} y^4 \qquad (12)$$

In reference [4] the author did experiments in brain imaging and image feature extraction using the above nonlinearities, however, no one had researched the role of nonlinearities in speech feature extraction. The paper would discuss the selection of nonlinearities and their coefficients. When a is equal to 1, $G_{opt}(y) = |y|$ is not derivable in origin and always replaced by $G(y) = \log(\cosh(a1 * y))$, where a1 is a real number. The two functions are as Fig. 2:

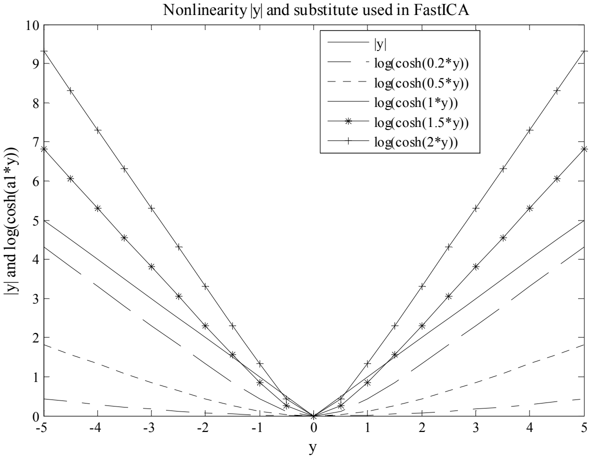


Figure 2. Gopt (y) = |y| and G(y) = log (cosh(a1*y))

As can be seen from Fig. 2, the closer to 1 the coefficient a1 is, the closer to optimal function |y| the function $\log(\cosh(a1 * y))$ is, while the curve is steeper. Reference [12] proposed that G(y) should not grow too fast with |y|, otherwise it would rely on some observations that are far from the origin. We must make a compromise between the approximation of accuracy and the function's smoothness. The selection of a1 will rely to specific applications and is not absolute. When using $\log(\cosh(a1 * y))$ as the function to extract speech features, we should test a1 and find the optimal value.

D. Features Selection

As for DCT, the reserved first several coefficients can be treated as speech features[13]. For ICA, feature selection can be processed according to the L2-norm of ICA basis. ICA basis refers to the column vector of the inverse of ICA demixed matrix. The L2-norm of basis represents the contribution of the basis to the whole signal. The bigger the value of the norm is, the more the contribution is. The L2-norm of ICA bases are shown in Fig. 3.

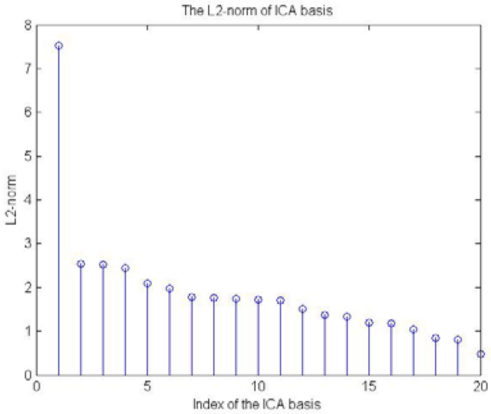


Figure 3. The L2 norm of bases of ICA

Therefore the most important N basis can be chosen according to the bigger value of L2-norm value of bases and used in speech feature recognition.

III. Experiments

A. Results of Four Kinds of Features

The paper used speech recognition development toolkit HTK of Cambridge University to build a speech recognition system based on HMM which was used to assess the new features based on symmetric orthogonalization ICA and other features . Moreover, the paper compared the new features with MFCC, PCA features, and the features based on deflation orthogonalization ICA. The experiments were performed on the Aurora-2 corpus, sampled to 8 kHz. The 80 speakers (consisting of 40 male and 40 female) from the training subset were selected. The training set consisted of 100 male and 100 female speeches which were clean. There were 40 male and 40 female speeches in each noise environment and SNR in testing set. There were 8 kinds of practical environments: airport, babble, car exhibition, restaurant, street, subway and train. The SNR were divided into 7 classes: -5db, 0db, 5db, 10db, 15db, 20db and clean. The total number of training and testing set was 4680.

The speech signal was divided into frames of 32 ms in length with an overlap of 10ms between frames. Pre-emphasis and Hamming window were applied to each frame first. Then FFT was performed then to extract the spectrum. Mel filter bank analysis with 20 channels was processed for each frame. Logarithm operation was performed following FFT. These coefficients were transformed by the DCT to get the traditional MFCC features. By using the PCA-based, deflation ICA-based and symmetric ICA-based transformation, PCA features and ICA features were obtained. The last two ICA-based features were denoted as ICA_DEFL and ICA_SYMM respectively. The final feature vector consisted of 13 components with first-order deltas and second-order deltas. There were 39 components in the final features in each case. The experiments extracted the four features in Matlab and tested them in a speech recognition system built using HTK. The results were as Fig. 4:

As can be seen from Fig. 4, when SNR was high or in clean condition, the performance of ICA_SYMM was almost the same as the other three features, while features based on ICA were superior to others in low SNR. Table 1 and Fig. 5 showed the average recognition rate of the four features in 8 noise environments. From them two, we can find the two features based on ICA were both superior to MFCC greatly. The recognition rate of ICA_SYMM feature was about 6.17% higher to MFCC. At the same time, ICA-based features were better than PCA-based feature, excepting that ICA_DEFL was lower than PCA in exhibition environment. The ICA_SYMM feature was better than ICA_DEFL feature, proving the accuracy of discussion of the two orthogonalization of ICA.

B. The Nonlinearities and Their Coefficients

In FastICA, the common nonlinearities of objective functions were (10), (11) and (12). Reference [14] proposed using rational nonlinearities as substitutes of the three common nonlinearities to reduce the mass of computation. However, experiments proved that it didn't work well. In our experiments, ICA_SYMM features were computed using the three nonlinearities respectively. The results were as Table 2 and Fig. 6.

As can be seen from Fig. 6, $G_1(y)$ was superior to $G_2(y)$ and $G_3(y)$ when computing average speech recognition rate in all 8 noise environments.

The coefficient a1 would affect the property of $G_1(y)=\frac{1}{a1}\log(\cosh(a1*y))$. The experiments researched the influence of a1 on recognition rate when its value was between 0 and 2. Results of the experiments were as Table 3 : in the street, car, airport, exhibition, restaurant and train environment, the maximum average recognition rate was reached when a1 equaled 0.2; in babble environment, the maximum average recognition rate 67.32% was reached when a1 equaled 1.1, but it was only 0.65% higher than that gotten when a1 equaled 0.2; in subway environment, the maximum average recognition rate 66.03% was reached when a1 equaled 0.6, and it was 1.87% higher than that gotten when a1 equaled 0.6. As can be seen, when a1 equaled 0.2, excellent performance could be gotten in most noise environments, and the value performed quite well in other special environments too. The experiments proved that the value 0.2 of a1 is a reliable empirical value in most situations. The value 0.2 of a1 could be used to improve the performance of automatic speech recognition systems when extracting speech features using ICA-based method to extract speech features in in noise environments.

In Fig. 7 below: the horizontal axis was the value of a1, from 0.1 to 2, and the interval was 0.1; the vertical axis was the average recognition rate in noise environments. In (a) and (b), the value of a1 influenced the average speech recognition rate in a certain trend, and the best performance was reached when the value was 0.2.

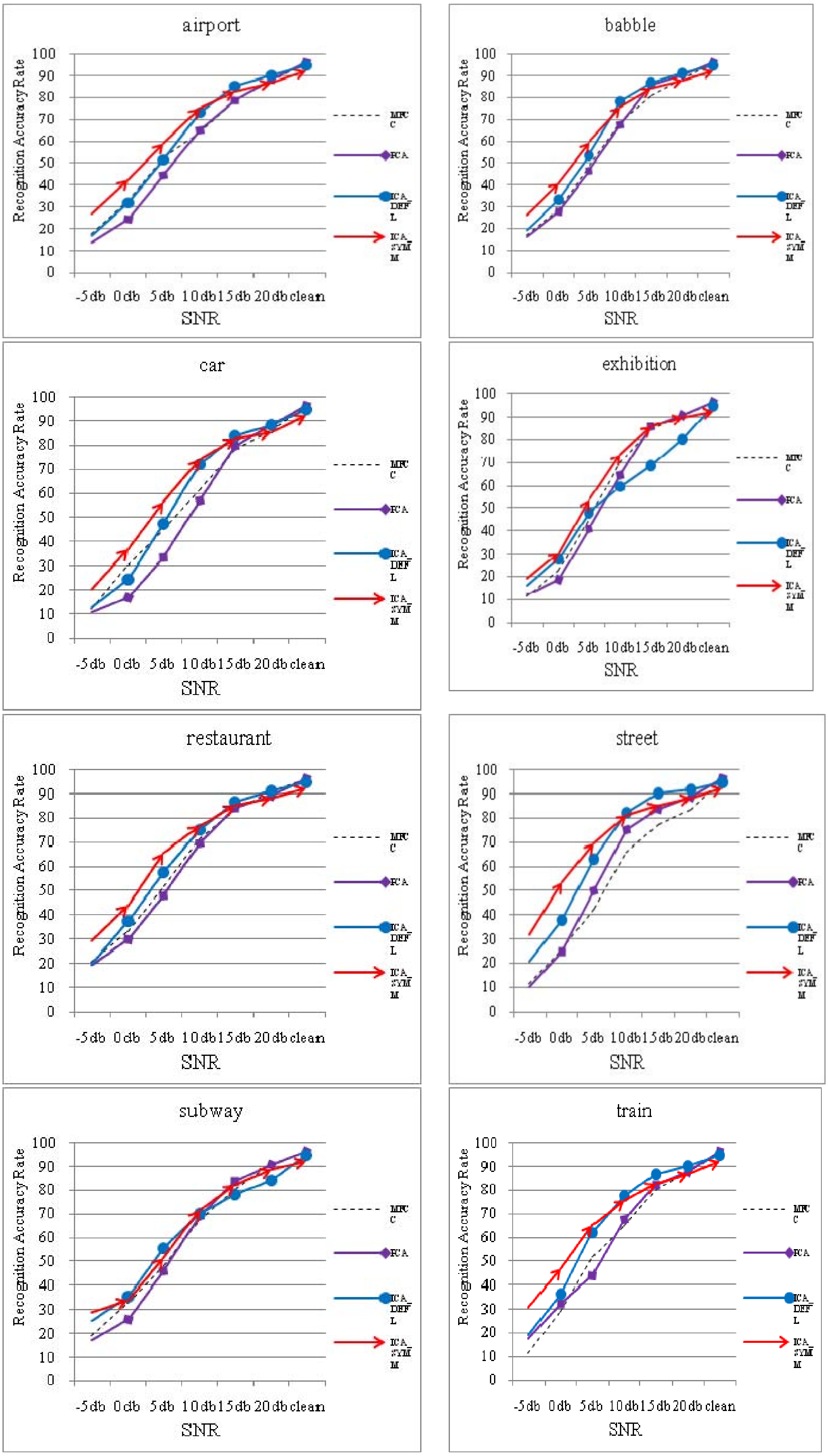


Figure 4. Recognition accuracy rate in each environment and each SNR (%)

TABLE I. AVERAGE RECOGNITION RATE OF THE FOUR FEATURES IN 8 ENVIRONMENTS (%)

	street	**babble**	**car**	**train**	**subway**	**restaurant**	**airport**	**exhibition**
MFCC	57.61	61.18	58.41	60.17	62.14	64.01	61.34	60.06
PCA	61.29	61.51	54.63	61.03	61.29	62.40	58.73	58.46
ICA_DEFL	68.59	65.18	60.54	66.72	63.42	66.14	63.37	56.49
ICA_SYMM	**71.78**	**66.67**	**64.11**	**68.69**	**64.16**	**68.80**	**66.51**	**63.57**

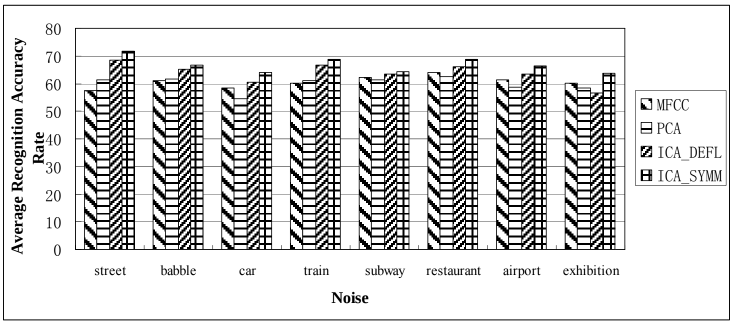


Figure 5. Average recognition rate of the four features in 8 environments (%)

TABLE II. AVERAGE RECOGNITION RATE OF THE THREE NONLINEARITIES IN 8 ENVIRONMENTS (%)

	airport	**babble**	**car**	**exhibition**	**restaurant**	**street**	**subway**	**train**
$G_2(y)$	56.25	57.63	51.56	54.38	58.49	60.62	54.28	55.29
$G_3(y)$	58.02	60.41	54.60	56.36	58.87	66.44	58.12	60.36
$G_1(y)$	**66.51**	**66.67**	**64.11**	**63.57**	**68.80**	**71.78**	**64.16**	**68.69**

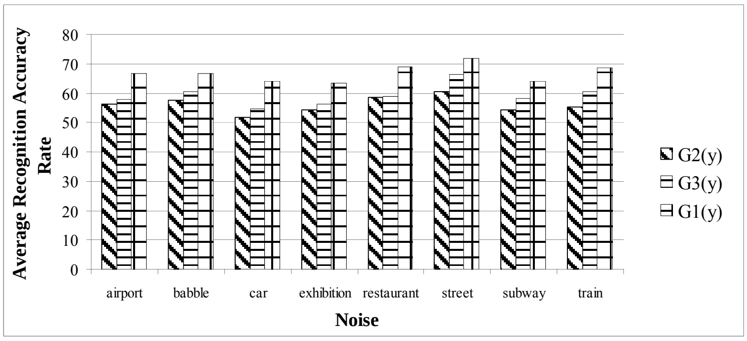
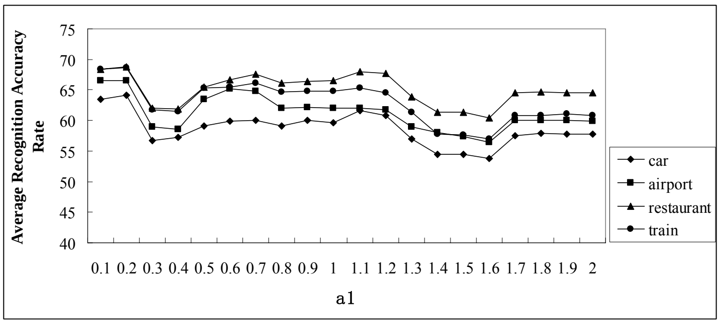
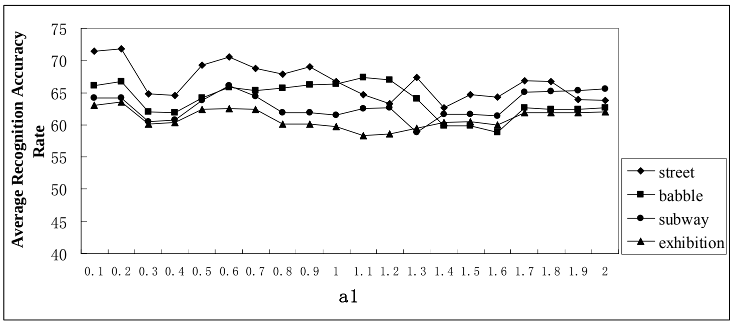


Figure 6. Average recognition rate of the three nonlinearities in 8 environments (%)



(a)



(b)

Figure 7. Average recognition rate influenced by different value of a1 in 8 environments (%)

TABLE III. THE AVERAGE RECOGNITION RATE IN DIFFERENT VALUES OF A1 IN 8 ENVIRONMENTS (%)

	street	**babble**	**car**	**subway**	**airport**	**exhibition**	**restaurant**	**train**
0.1	71.46	66.13	63.47	64.21	66.51	63.04	68.37	68.32
0.2	**71.78**	66.67	**64.11**	64.16	**66.51**	**63.57**	**68.80**	**68.69**
0.3	64.85	61.97	56.75	60.54	58.97	60.06	61.98	61.71
0.4	64.58	61.87	57.18	60.75	58.56	60.32	61.87	61.49
0.5	69.23	64.22	59.16	63.80	63.53	62.36	65.39	65.29
0.6	70.51	65.82	59.85	**66.03**	65.18	62.57	66.62	65.50
0.7	68.74	65.28	60.01	64.38	64.75	62.35	67.58	66.14
0.8	67.88	65.76	59.10	61.92	62.03	60.11	66.14	64.70
0.9	69.01	66.19	60.01	61.93	62.19	60.12	66.41	64.86
1.0	66.72	66.30	59.69	61.55	61.98	59.74	66.56	64.86
1.1	64.75	**67.32**	61.66	62.56	62.03	58.35	68.00	65.28
1.2	63.35	66.99	60.86	62.67	61.76	58.56	67.68	64.53
1.3	67.42	64.06	56.97	58.78	59.00	59.47	63.90	61.34
1.4	62.62	59.89	54.46	61.66	57.98	60.37	61.28	57.71
1.5	64.75	59.84	54.51	61.66	57.39	60.53	61.34	57.60
1.6	64.37	58.88	53.82	61.33	56.48	59.95	60.37	56.96
1.7	66.88	62.67	57.55	65.02	60.06	61.87	64.48	60.75
1.8	66.72	62.41	57.87	65.18	60.01	61.87	64.64	60.86
1.9	63.95	62.46	57.82	65.34	60.06	61.88	64.54	61.02
2.0	63.84	62.61	57.82	65.60	59.95	61.98	64.54	60.86

IV. CONCLUSION

The paper proposed using symmetric orthogonalization ICA-based method to extract speech features, and verified the new features in 8 different kinds of noise environments. The experiments proved that the average recognition rate of the new features was 6.17% higher than that of MFCC features, especially excellent in low SNRs. The new method got better performance than deflation orthogonalization ICA-based method and MFCC, and improved the robustness of the speech recognition system and the efficiency of estimation of ICA. The nonlinearities of objective function in ICA and their coefficients had a great impact on recognition accuracy rate, when $G_1(y)=\frac{1}{a1}\log(\cosh(a1*y))$ and a1 = 0.2 it got the best performance in general. Because the demixed matrix of ICA in the new method was calculated offline, it could be calculated firstly before used in extracting speech features in fact which would save much time for estimation. From the above, we can see that the new method improved the average recognition rate but didn't strength the complexity of computation, so it is possible for the new method to replace MFCC as a popular method extracting speech features in the future.

REFERENCES

[1] U. Shrawankar and V. Thakare. "Feature Extraction for a Speech Recognition System in Noisy Environment: A Study," Computer Engineering and Applications (ICCEA), 2010 Second International Conference on. 2010.

[2] Z. Tuske, P. Golik, R. Schluter, and F.R. Drepper. "Non-stationary feature extraction for automatic speech recognition," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. 2011.

[3] W. Qiang, Z. Liqing, and S. Guangchuan, "Robust Multifactor Speech Feature Extraction Based on Gabor Analysis," Audio, Speech, and Language Processing, IEEE Transactions on, 2011. vol 19(4), pp. 927-936, 2011.

[4] H. Hsieh, J. Chien, K. Shinoda, and S. Furui. "Independent component analysis for noisy speech recognition," Acoustics, Speech and Signal Processing, (ICASSP). IEEE International Conference on. 2009.

[5] E. Ollila, "The Deflation-Based FastICA Estimator: Statistical Analysis Revisited," Signal Processing, IEEE Transactions on, 2010. vol **58**(3), pp. 1527-1541, 2011.

[6] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," Neural Networks, IEEE Transactions on, 1999. vol **10**(3), pp. 626-634, 1999.

[7] X. Zou, P. Jancovic, and M. Kokuer, "On the Effectiveness of the ICA-based signal representation in non-Gaussian Noise," Icsp: 2008 9th International Conference on Signal Processing, vols 1-5, pp. 1-4, 2008.

[8] P. Somervuo, "Experiments with linear and nonlinear feature transformations in HMM based phone recognition," 2003 Ieee International Conference on Acoustics, Speech, and Signal Processing, vol I, pp. 52-55, 2003.

[9] P. Hyunsin, T. Takiguchi, and Y. Ariki. "Integration of Phoneme-Subspaces Using ICA for Speech Feature Extraction and Recognition," Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008. 2008.

[10] R. Schluter, A. Zolnay, and H. Ney. "Feature combination using linear discriminant analysis and its pitfalls," INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, September 17, 2006 - September 21, 2006. 2006. Pittsburgh, PA, United states: DUMMY PUBID.

[11] L. Potamitis, N. Fakotakis, and G. Kokkinakis, "Independent component analysis applied to feature extraction for robust automatic speech recognition," Electronics Letters, vol **36**(23) pp. 1977-1978, 2000.

[12] A. Hyvärinen, J. Karhunen, and E. Oja, Independent component analysis. vol. 26, 2001.

[13] J.H. Lee, H.Y. Jung, T.W. Lee, and S.Y. Lee, "Speech feature extraction using independent component analysis," 2000 Ieee International Conference on Acoustics, Speech, and Signal Processing, vols I-Vi, pp.1631-1634, 2000.

[14] P. Tichavský, Z. Koldovský, and E. Oja, "Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions," Independent Component Analysis and Signal Separation, pp. 285-292, 2007.

Huan Zhao is a professor at the School of Information Science and Engineering, Hunan University. She obtained her B.Sc. degree, M.S. degree and Ph.D. in Computer Science and Technology at Hunan University in 1989, 2004 and 2010, respectively. Her current research interests include speech information processing, embedded system design and embedded speech recognition. She served as visiting scholar at the University of California, San Diego (UCSD), USA during the period of March 2008 to September 2008. The visiting scholarship was appointed and sponsored by the China Scholarship Council (CSC). Prof. Zhao is a Senior Member of China Computer Federation, Governing of Hunan Computer Society, China and China Education Ministry Steering Committee Member of Computer Education on Arts. She has published more than 40 papers and 9 books.

Kai Zhao received his B.Sc. degree in Computer Science and Technology at the school of Computer and Communication, Hunan University, P. R. China in 2009. Currently, he is a M.S. candidate of Hunan University, P. R. China. His current research interests include speech feature extraction and speech recognition.

He Liu received her B.Sc. degree in Electronic Information Engineering at the School of Computer and Electronic Engineering, Hunan University of Commerce, P. R. China in 2009. Currently, she is a M.S. candidate of Hunan University, P. R. China. Her current research interests include digital signal processing, speech information processing and feature extraction.