A FISHERVOICE BASED FEATURE FUSION METHOD FOR SHORT UTTERANCE SPEAKER RECOGNITION

Chenhao Zhang
2013/04/08

Outline

- Introduction
- The FisherVoice based SUSR framework
- Experimental Results and Analysis
- Conclusions

Introduction

- Short Utterance Speaker Recognition (SUSR)
  - In some situations, only a short utterance containing only one or two words is available
  - Short utterances can provide a better user experience
  - The current technologies are unsatisfactory when the test speech is very short
- GMM-UBM / GMM-SVM
  - Dominant speaker recognition Technologies
  - The classic and effective methods when the test data is enough
  - The performance degrades sharply when the test speech is shortened

The Influence of the Length of Test Speech

- The length of the test data is a big factor that influences the performance of speaker recognition
  - R. Vogt, S. Sridharan and Michael Mason. IEEE Trans on ASLP 2010. On NIST SRE 2005 Database

Valid length	**EER(%)**	**MinDCF (10^{-2})**
Sufficient	6.34	2.93
20 Seconds	8.87	3.91
10 Seconds	12.15	4.89
5 Seconds	16.99	6.16
2 Seconds	23.89	7.94
< 2 Seconds	>35	>10

Existing Solutions

- Factor Analysis Subspace Estimation
  - Decrease the number of redundant model parameters to develop dominant speaker models [P. Kenny 2005]
- Speech Segments Selection
  - Select segments with higher discriminability on speaker characteristics [M. Nosratighods 2010]
- Score Fusion
  - Weighted bilateral scoring [A. Malegaonkar 2008]
- Most of the above mentioned approaches show improvements with test length among 5~10 seconds.

Feature Combination

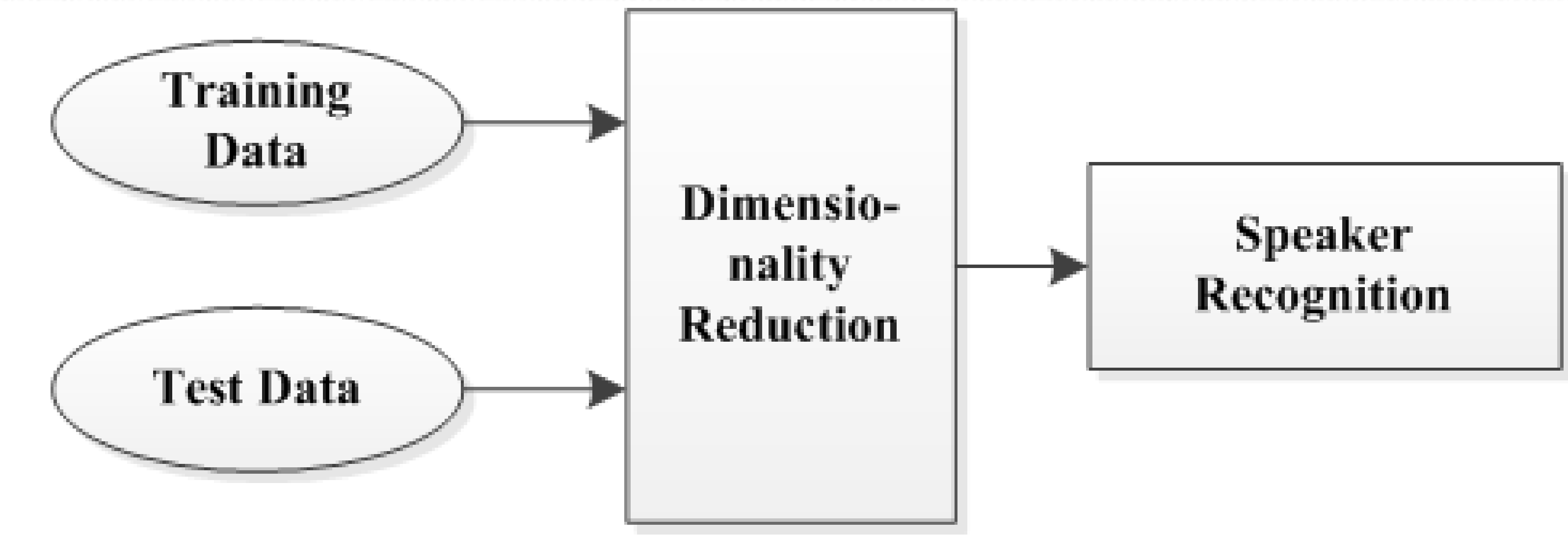
- The same speaker recognition system with different kinds of features will perform quite differently
  - Mel Frequency Cepstral Coefficients (MFCC)
- For short utterance, the information of one kind feature will not be enough
  - One single kind of feature can provide relatively enough speaker information to perform speaker recognition when the test utterance is long enough
- The combination of different features is useful to improve the recognition performance in many research fields
  - Feature Fusion. [J. Yang, 2003]

Feature Fusion Method

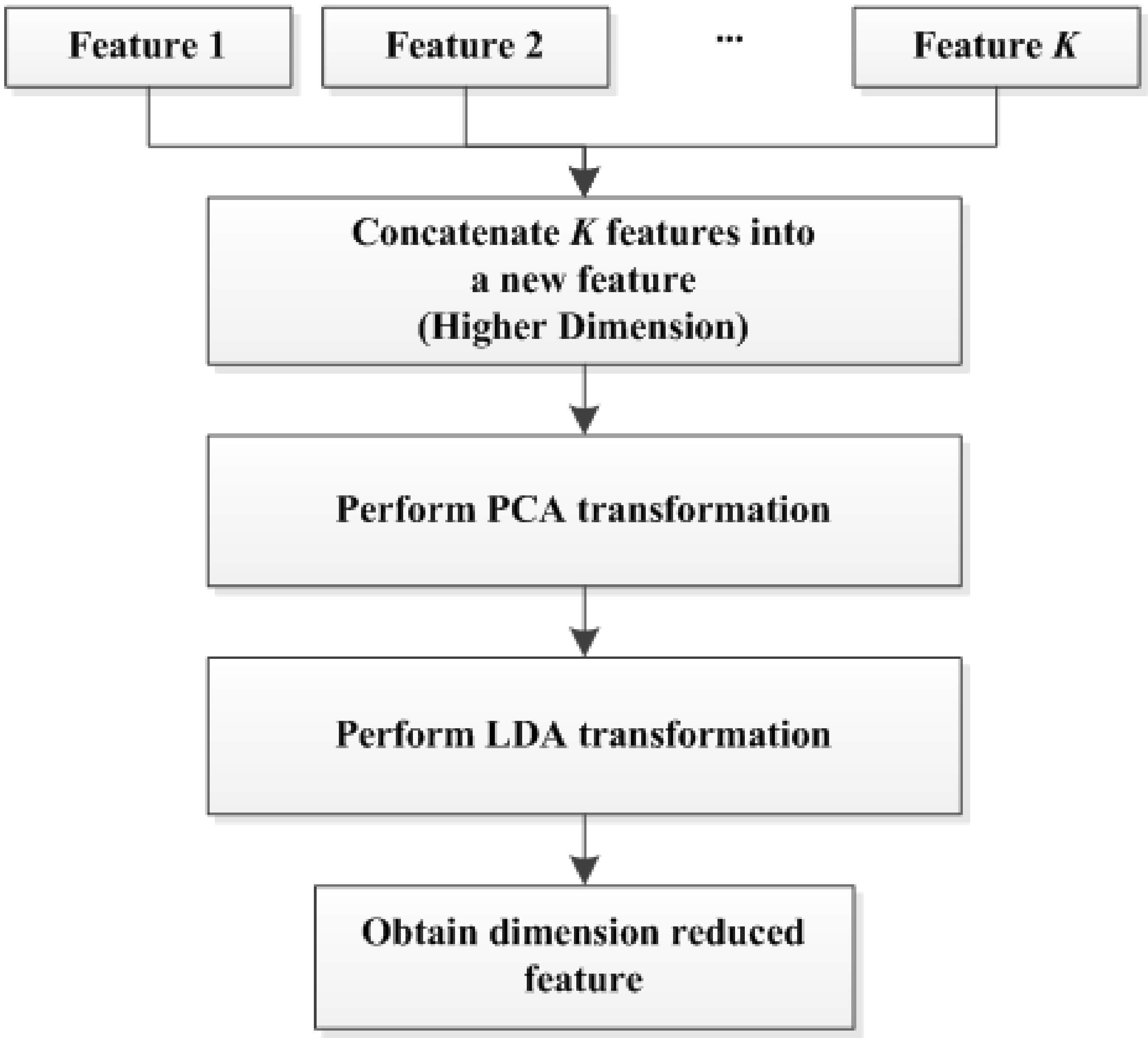
- Target: 2 aspects
  1. De-correlate the concatenated feature vectors into individual ones from multiple feature streams
  2. Eliminate the coefficients with redundant and unimportant information
- Linear Discriminant Analysis (LDA)
  - Maximize the between-class covariance and simultaneously minimizing the within-class covariance
  - Problem: The Singular Matrix
- The Fishervoice based method
  - Principal Component Analysis (PCA) plus LDA

The FisherVoice based SUSR Framework

- Two Key Parts
  1. The Fishervoice based dimensionality reduction
     - combine different kinds of features
  2. The GMM-UBM based speaker recognition



The Fishervoice based Dimensionality Reduction



Linear Discriminant Analysis

- For the original data set **X**, the within-class scatter matrix and the between-class scatter matrix are:

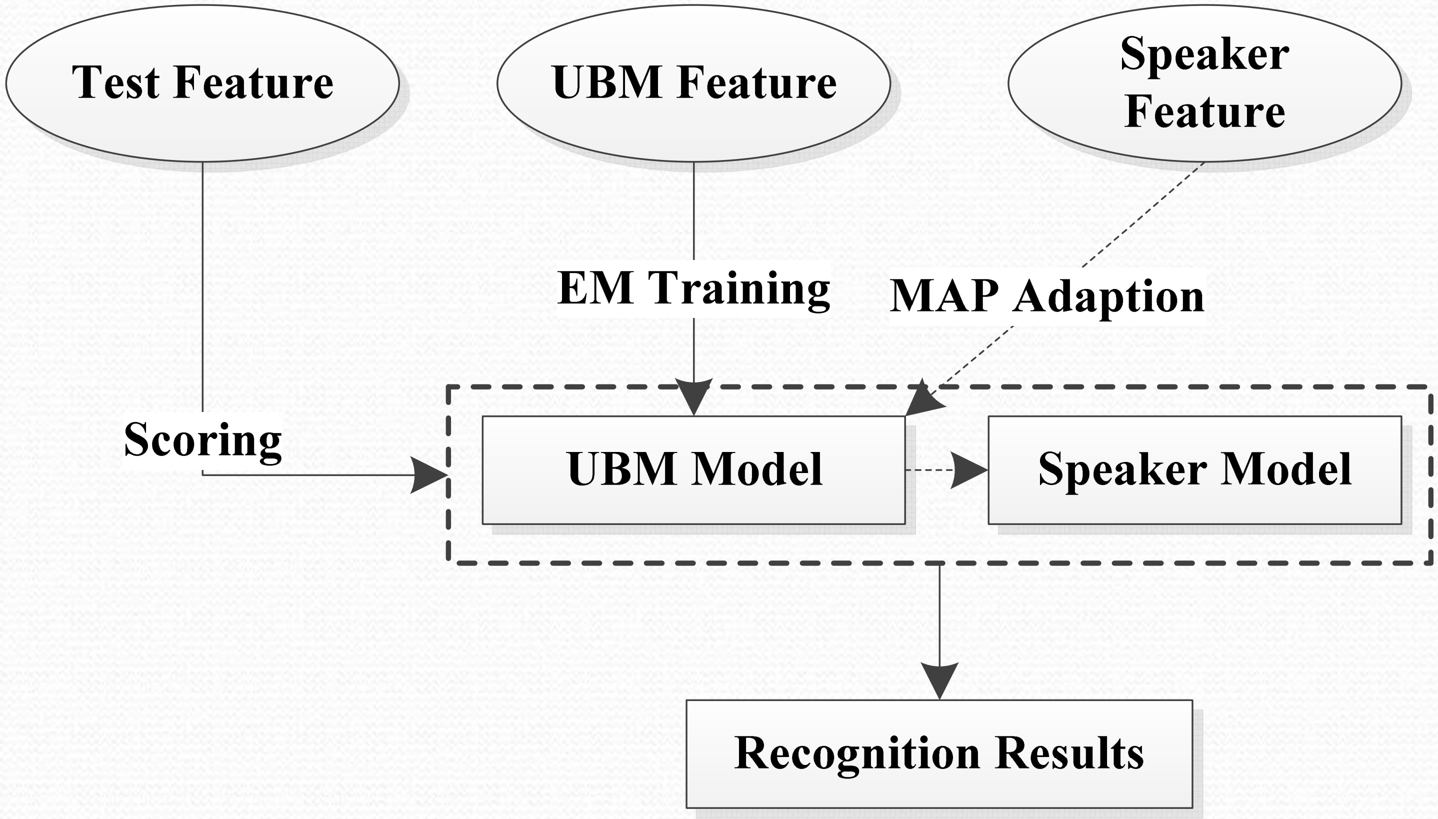
$$S_W = \sum_{c=1}^{C} \sum_{\mathbf{x} \in D_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T \qquad S_B = \sum_{c=1}^{C} n_c (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$$

$$\mathbf{m}_c = \frac{1}{n_c} \sum_{\mathbf{x} \in D_c} \mathbf{x} \qquad \mathbf{m} = \frac{1}{\sum_{c=1}^{C} n_c} \sum_{c=1}^{C} n_c \mathbf{m}_c$$

- LDA maximizes the criterion as:

$$J(W) = \frac{\tilde{S}_B}{\tilde{S}_W} = \frac{W^T S_B W}{W^T S_W W}$$

The GMM-UBM based speaker recognition



Database

- Database: SUD12
  - 60 Chinese speakers: 30 males and 30 females
  - 163 Chinese sentences:
    - 100 long sentences for train / 63 short sentences for test
  - The Distribution of the length of the test utterances

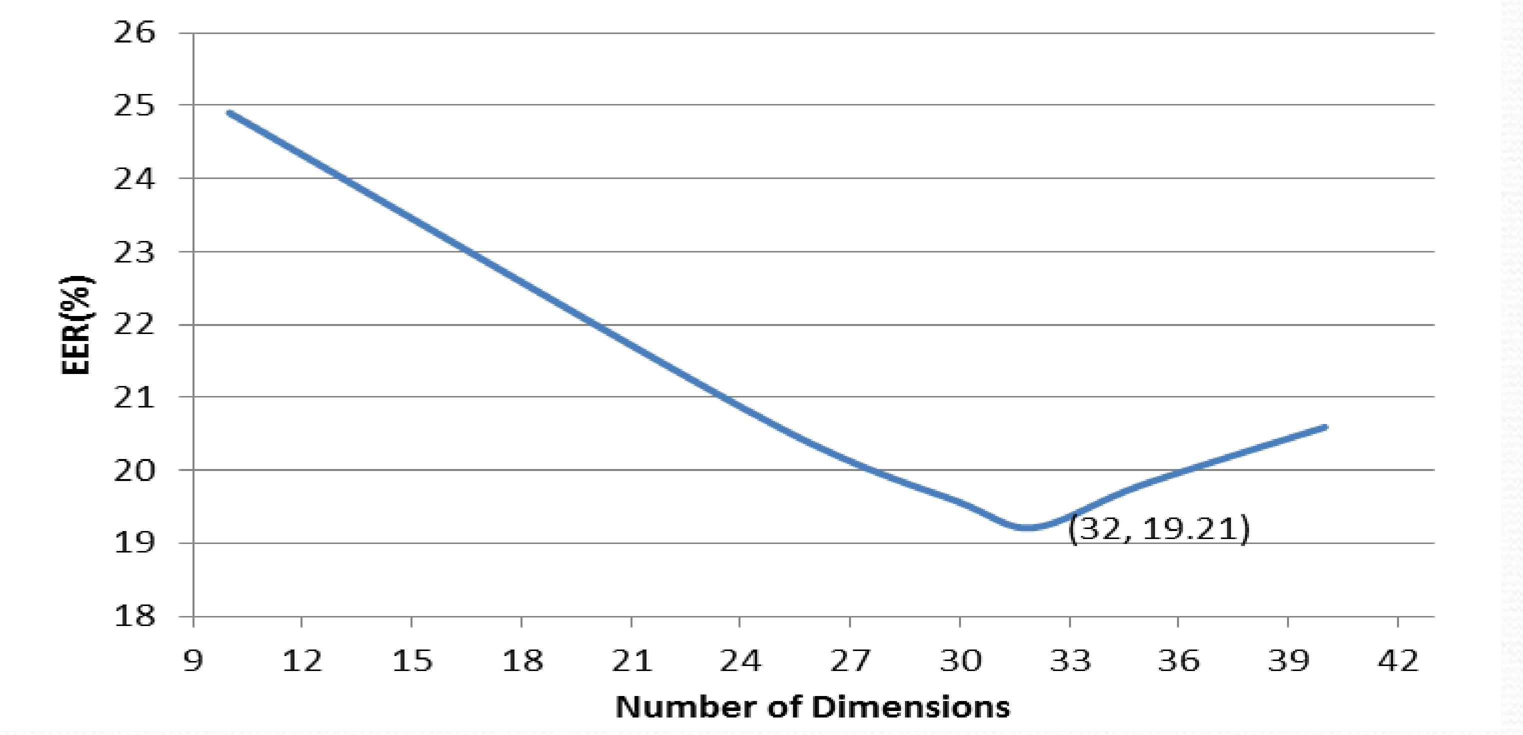
Length in second	# of Sentences	Percent (%)
less than 0.5	38	60.32
0.5 to 1.0	15	23.81
1.0 to 2.0	10	15.87

  - Recorded in clean environments using a microphone at 8 kHz sampling rate with 8-bit precision

Experimental Conditions

- Three kinds of features were and they are:
  - **MFCC** - 20-dimensional Mel Frequency Cepstral Coefficients (MFCC)
    - 30 Mel filter banks.
  - **PLAR** - 20-dimensional Perceptual Log Area Ratio (PLAR) Be robust to the noise and other environments [19], is derived from the Perceptual Linear Prediction feature (PLP) [20].
  - **LPCC** - 12-dimensional Linear Predictive Cepstrum Coefficients (LPCC)
- 52-dimensional feature vector

Results and Analysis



EER of the Fishervoice based method as a function of number of dimensions

Results and Analysis

Feature	EER (%)
MFCC	26.52
PLAR	22.98
LPCC	23.44
Concatenated Feature	28.78
LDA based	20.03
Fishervoice based	19.21

Conclusions

- The feature fusion method can improve the performance when the test utterance is short.
- The proposed Fishervoice based method can achieve a better result compared with the traditional features and the LDA fusion method in short test utterance situations.
- The feature domain method can be combined with methods from other domains to achieve a better performance for SUSR

References

1. J. P. Campbell. Speaker recognition: a tutorial. Proceedings of the IEEE, 1997, vol. 85, pp. 1437-1462
2. D. A. Reynolds, T. Quatieri, R. Dunn. Speaker verification using adapted gaussian mixture models. Digital Signal Processing, 2000, vol. 10, pp. 19-41
3. W. M. Campbell et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP, 2006, pp. 97-100
4. R. Vogt, S. Sridharan and Michael Mason. Making confident speaker verification decisions with minimal speech. IEEE Trans. on ASLP, 2010, vol. 18, no. 6, pp. 1182-1192
5. NIST Speaker Recognition Evaluation Plan, Online Available http://www.nist.gov/speech/tests/sre/
6. P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. IEEE Trans. on Speech and Audio Processing, 2005, vol. 13, no. 3, pp. 345-354
7. N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny. Cosine similarity scoring without score normalization techniques. Odyssey 2010
8. M. Nosratighods, E. Ambikairajah, J. Epps and M. J. Carey. A segment selection technique for speaker verification. Speech Communication, 2010, pp. 753-761
9. A. Malegaonkar, A. Ariyaeeinia. On the enhancement of speaker identification accuracy using weighted bilateral scoring. ICCST, 2008
10. S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on ASLP, 1980, vol. 28, no. 4, pp. 357–366

References

11. D. Chow and W. H. Abdulla. Robust speaker identification based perceptual log area ratio and Gaussian mixture models. Interspeech , 2004
12. P. Premakanthan and W. B. Mikhael. Speaker verification /recognition and the importance of selective feature extraction: review. MWSCAS, 2001, vol. 1, 57-61.
13. J. Yang, J.-Y Yang, D.Zhang and J.-F Lu. Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition, 2003, vol. 336, no.6, pp.1369-1381
14. D. Cai, X.-F He, J.-W Han. SRDA: An efficient algorithm for large-scale discriminant analysis. IEEE Trans on Knowledge and Data Engineering 2008, vol. 20, No. 1
15. P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE TPAMI, 1997, vol. 19, no. 7
16. I. Joliffe. Principal component analysis. Springer-Verlag, 1986
17. R. Duda, P. Hart, and D.G. Stork. Pattern Classification. 2nd edition, 2001
18. J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 1998
19. C.-H Zhang, X.-J Wu and T.F. Zheng. A K-Phoneme-Class based multi-model method for short utterance speaker recognition. APSIPA, 2012
20. D. Wang, X. Zhu, Y. Liu. Multi-layer channel normalization for frequency-dynamic feature extraction. Journal of Software, 2005, vol. 12, no. 9, pp.1523-1529
21. Z. Bai and H. Zha. A new preprocessing algorithm for the computation of the generalized singular value decomposition. SIAM Journal on Scientific Computing, 1993

Thank you!