FISHERVOICE AND SEMI-SUPERVISED SPEAKER CLUSTERING

Stephen M. Chu[1]*, Hao Tang*[2]*, Thomas S. Huang*[2]

[1]IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598, USA
[2]Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, I.L. 61801, USA

ABSTRACT

Speaker subspace modeling has become increasingly important in speaker recognition, diarization, and clustering. Principal component analysis (PCA) is a popular linear subspace learning technique and the approach that represents an arbitrary utterance or speaker as a linear combination of a set of basis voices based on PCA is known as the eigenvoice approach. In this paper, a novel technique, namely the fishervoice approach, is proposed. The fishervoice approach is based on linear discriminant analysis, another successful linear subspace learning technique that provides an optimized low-dimensional representation of utterances or speakers with focus on the most discriminative basis voices. We apply the fishervoice approach to speaker clustering in a semi-supervised manner and show that the fishervoice approach significantly outperforms the eigenvoice approach in all our experiments on the GALE Mandarin dataset.

Index Terms— Linear subspace learning, eigenvoice, fishervoice, semi-supervised speaker clustering.

1. INTRODUCTION

Traditional techniques for speaker recognition [1], diarization [2, 3, 4], and clustering [5, 6, 7, 8, 9, 10, 11] are primarily based on modeling the probability distributions of low-level acoustic features [12, 13] and decision making with statistical distance metrics such as the log likelihood ratio [14]. The probability distributions are typically modeled by the Gaussian mixture models (GMM) [14] and the standard training method is to adapt the mixture means, covariance matrices and weights of a universal background model (UBM) [15], trained over a considerable amount of speech data from a set of speakers, to a small amount of speech data (e.g., an utterance) from a target speaker by *maximum a posterior* (MAP) adaptation [15].

Recent studies on speaker and channel variability for speaker recognition have led to an idea of stacking the mixture means of a GMM to form a GMM mean supervector [16]. In this way a given utterance or speaker can be represented as a single point in a high dimensional space - the speaker space. Alternatively, a speaker can be represented as a cluster of points in the speaker space. This utterance/speaker representation reduces the speaker recognition, diarization, and clustering problems to those of general pattern recognition where many successful pattern recognition techniques such as nearest neighbor (NN) and support vector machine (SVM) as well as many useful distance metrics such as the Euclidean or Mahalanobis distance can be naturally employed.

Pattern recognition in high dimensional spaces has inherently led to many significant problems, noticeably the "curse of dimensionality" [17, 18]. Most often, the data points in a high dimensional space do not usually scatter uniformly. Rather, they lie in or near a low-dimensional subspace (i.e., a low-dimensional manifold) of the original space. Thus, it is always good practice that we perform some sort of dimensionality reduction or subspace learning before a particular pattern recognition algorithm is applied. Principal component analysis (PCA) [19] and linear discriminant analysis (LDA) [20] are two popular linear subspace learning techniques which have been very successful in their respective application scenarios. PCA is an unsupervised technique that aims to find a low-dimensional subspace while maximizing the variances reserved, and thus is optimal for data reconstruction. LDA is a supervised technique that aims to find a low-dimensional subspace while maximizing the separability of the individual classes, and thus is optimal for data classification.

In the literature, speaker subspace learning has begun to gain its popularity and the approach that represents an arbitrary utterance or speaker as a linear combination of a set of basis voices based on PCA is known as the eigenvoice approach [21, 22]. The eigenvoice approach has been demonstrated to be successful for both speaker recognition [23] and diarization [24], and to outperform algorithms directly working in the original speaker space. However, since an important goal of pattern recognition is classification, it is natural that we prefer a subspace learning technique that is discriminative rather than generative, which is beneficial to classification. In this paper, we propose a novel technique, namely the fishervoice approach. The fishervoice approach is a speaker subspace learning technique based on LDA. The term "fishervoice" is analogous to "fisherface" in the face recognition literature, where the fisherface approach is the face recognition method based on LDA while the eigenface approach is based on PCA [20].

In this paper, we performed semi-supervised speaker clustering experiments on the GALE Mandarin dataset [25]. Here, semi-supervision refers to the fact that we use an independent training data set to assist the unsupervised speaker clustering process. In the experiments, we applied both the eigenvoice and fishervoice approaches. Our experiment results clearly indicate that the fishervoice approach significantly outperforms the eigenvoice approach in all the experiments, while the eigenvoice approach slightly outperforms direct speaker clustering in the original speaker space. The effectiveness of the fishervoice is rooted in the fact that it is based on discriminant analysis whose goal is to facilitate classification - a generalized component of speaker recognition, diarization, and clustering.

2. MAP ADAPTATION AND GMM MEAN SUPERVECTOR

The low-level acoustic features of an utterance is modeled by an M-mixture GMM, which is defined as a weighted sum of M component

This work was funded in part by DARPA contract HR0011-06-2-0001.

Gaussian densities

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i N(\mathbf{x}|\mu_i, \Sigma_i) \quad (1)$$

where $\mathbf{x}$ is a D-dimensional feature vector, w_i is the i^{th} mixture weight satisfying the constraint $\sum_{i=1}^{M} w_i = 1$, and $N(\mathbf{x}|\mu_i, \Sigma_i)$ is a multivariate Gaussian probability density function

$$N(\mathbf{x}|\mu_i, \Sigma_i) = \frac{exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \quad (2)$$

with mean vector μ_i and covariance matrix Σ_i. The parameters of a GMM $\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{M}$ can be learned through the well-known expectation-maximization (EM) algorithm [26].

Usually, a single speaker-independent UBM, λ_0, is trained over all the utterances from all speakers in a training set, and for a particular utterance from a particular speaker, a GMM, λ, is derived by updating the well-trained UBM with that utterance via MAP adaptation. The MAP adaptation proceeds as follows. Starting with the UBM λ_0 and a training utterance $X = \{\mathbf{x}_t\}_{t=1}^{T}$, we first determine the probabilistic alignment of the training feature vectors into the UBM's component densities via Bayes rule

$$p(i|\mathbf{x}_t) = \frac{w_i N(\mathbf{x}_t|\mu_{0i}, \Sigma_{0i})}{\sum_{j=1}^{M} w_j N(\mathbf{x}_t|\mu_{0j}, \Sigma_{0j})} \quad (3)$$

Note Equation 3 is the posterior probability of assigning a training feature vector $\mathbf{x}_t$ to the i^{th} mixture component given the UBM λ_0. Next, we compute the following sufficient statistics

$$n_i = \sum_{t=1}^{T} p(i|\mathbf{x}_t) \quad (4)$$

$$E_i = \frac{1}{n_i}\sum_{t=1}^{T} p(i|\mathbf{x}_t)\mathbf{x}_t \quad (5)$$

$$E_i^2 = \frac{1}{n_i}\sum_{t=1}^{T} p(i|\mathbf{x}_t)\mathbf{x}_t^2 \quad (6)$$

Finally, the above sufficient statistics are used to update the model parameters

$$w_i' = [\alpha_i n_i/T + (1-\alpha_i)w_i]\delta \quad (7)$$

$$\mu_i' = \beta_i E_i + (1-\beta_i)\mu_i \quad (8)$$

$$\sigma_i'^2 = \gamma_i E_i^2 + (1-\gamma_i)(\sigma_i^2 + \mu_i^2) - \mu_i'^2 \quad (9)$$

where δ is a scaling factor computed over all adapted mixture weights to ensure that they sum to unity. The adaptation coefficients used in Equations 7-9 are smartly given by the empirical formula

$$\nu_i = \frac{n_i}{n_i + r^\nu} \quad (10)$$

where $\nu \in \{\alpha, \beta, \gamma\}$ and r^ν is a fixed relevance factor for ν.

Typically, only the mixture means are updated. Once a GMM is obtained for an utterance, its mixture means are stacked to form a GMM mean supervector

$$\mathbf{x} = [\mu_1{}^T \quad \mu_2{}^T \quad ... \quad \mu_M{}^T]^T \quad (11)$$

It is numerically beneficial that we subtract from a GMM mean supervector the mean supervector of the UBM, i.e.,

$$\mathbf{x}' = \mathbf{x} - \mathbf{x}_0 \quad (12)$$

where $\mathbf{x}_0$ is the mean supervector of the UBM. In this context, without causing any ambiguity, instead of $\mathbf{x}$, the difference $\mathbf{x}'$ is called the GMM mean supervector. A complete set of GMM mean supervectors forms a high-dimensional space called the speaker space.

3. THE FISHERVOICE APPROACH

The fishervoice approach represents an arbitrary utterance or speaker as a linear combination of a set of basis voices, which are learned in a supervised manner through an independent training data set using LDA. Fisher's original LDA is proposed for the two-class case. However, it can be extended to deal with multiple classes [17]. In multi-class LDA, the data are projected from the original D-dimensional space to a $(C-1)$-dimensional subspace such that the within-class scatter of the projected data is minimized while the between-class scatter of the projected data is maximized. Assuming that a training set $D_1 \bigcup D_2 \bigcup ... \bigcup D_C$ contains training examples from C classes, the within-class scatter matrix of the data is

$$S_W = \sum_{c=1}^{C}\sum_{\mathbf{x}\in D_c}(\mathbf{x}-\mathbf{m_c})(\mathbf{x}-\mathbf{m_c})^T \quad (13)$$

where $\mathbf{m_c} = \frac{1}{n_c}\sum_{\mathbf{x}\in D_c}\mathbf{x}$ is the mean vector of the c^{th} class and $n_c = |D_c|$ is the number of examples in D_c. The between-class scatter matrix of the data is

$$S_B = \sum_{c=1}^{C} n_c(\mathbf{m_c}-\mathbf{m})(\mathbf{m_c}-\mathbf{m})^T \quad (14)$$

where $\mathbf{m} = \frac{1}{n}\sum_{c=1}^{C} n_c\mathbf{m_c}$ is the total mean vector of the data and $n = n_1 + ... + n_C$.

LDA projects the data from the D-dimensional space to a $(C-1)$-dimensional subspace through a linear transformation $Y = W^T X$ where the columns of X are the examples $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ in the original space and the columns of Y are the corresponding examples projected into the lower-dimensional subspace $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$.

The within-class and between-class scatters of the projected data set Y, $\widetilde{S}_W$ and $\widetilde{S}_B$, can be likewise computed in a way similar to Eqs. 13-14. It is straightforward to show that

$$\widetilde{S}_W = W^T S_W W, \quad \widetilde{S}_B = W^T S_B W \quad (15)$$

and LDA seeks to maximize the following criteron

$$J(W) = \frac{\widetilde{S}_B}{\widetilde{S}_W} = \frac{W^T S_B W}{W^T S_W W} \quad (16)$$

The solution to Eq. 16 can be obtained by solving a generalized eigenvalue problem

$$S_B\mathbf{w_i} = \lambda_i S_W\mathbf{w_i}, \quad i = 1, 2, ..., n \quad (17)$$

The columns of the $D \times (C-1)$ projection matrix W consist of the generalized eigenvectors $\mathbf{w_i}$ corresponding to the $C-1$ largest eigenvalues of $\{S_B, S_W\}$.

In practice, the within-class scatter matrix S_W is always singular with its rank being at most $n-C$. In order to overcome this difficulty, PCA is always first applied to reduce the dimensionality of the data vectors to $n-C$ or less to ensure the non-singularity of S_W before LDA is carried out.

4. SEMI-SUPERVISED SPEAKER CLUSTERING

Speaker clustering is the process of assigning individual utterances to their respective speakers and is an important part of the larger task of speaker diarization. Normally, speaker clustering is completely unsupervised. There is no training data available or required. In this paper, we propose a new speaker clustering concept, *semi-supervised speaker clustering*, in which semi-supervision refers to the use of our prior knowledge of speakers in general to assist the unsupervised speaker clustering process while speaker clustering is still performed in an unsupervised manner.

In semi-supervised speaker clustering, the prior knowledge of speakers is obtained via an independent training data set, which can help us learn the following:

- a speaker-discriminative feature transformation. This will be explained in detail in Section 5.
- a speaker-independent prior model - the UBM.
- a discriminative speaker subspace - the fishervoice space.

Fig. 4 shows the functional diagram of semi-supervised speaker clustering. First, a training set is used to train a UBM via the EM algorithm. The UBM is then MAP adapted to give a GMM for every training utterance. Any utterance in the training set can thus be represented by a GMM mean supervector. Next, LDA (or PCA) learning is performed on the training set in the GMM mean supervector space to derive a low-dimensional subspace - the fishervoice (or eigenvoice) space. Given a test utterance, the UBM is MAP adapted to form a GMM mean supervector for this test utterance, which is further projected onto the fishervoice (or eigenvoice) space. Finally, hierarchical agglomerative clustering (bottom-up) technique [27] based on the Euclidean distance metric and the Ward's linkage method [28] is used to perform unsupervised clustering in the fishervoice (or eigenvoice) space.

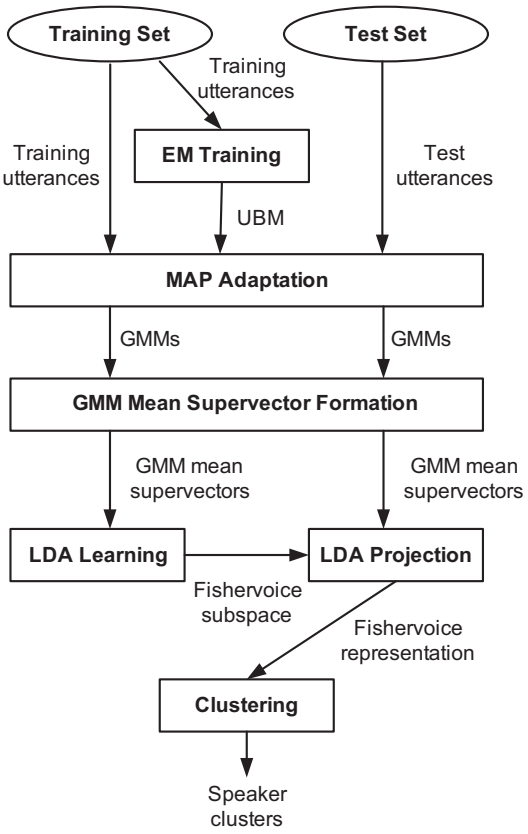

Fig. 1. Functional diagram of semi-supervised speaker clustering based on the fishervoice approach.

5. EXPERIMENTS

In this section, we apply both the eigenvoice and fishervoice approaches to the semi-supervised speaker clustering experiments conducted on the GALE Mandarin dataset [25] and present our experiment results. The GALE Mandarin dataset contains 1900 hours of broadcasting news speech data collected from various TV programs. The waveforms are sampled at 16K Hz and quantized at 16 bits per sample. We select a subset of the GALE Mandarin dataset that contains 498 speakers and 18324 utterances. On average, each speaker in this subset contains about 30-40 utterances, and each utterance is about 3-4 seconds long. This subset of speech data forms our independent training set.

The basic acoustic features we used are the 13 dimensional perceptual linear prediction (PLP) features [13], which are extracted using a hamming window of 25ms at a rate of 10ms per frame. Instead of computing the delta and delta-delta features we augment, for each frame, the basic PLP features with the PLP features of the neighboring frames. The PLP features of the current frame and those of the two frames to the left and those of the two frames to the right are concatenated to form a $13 \times 9 = 117$ dimensional long PLP feature vector. In this high-dimensional PLP feature vector space, LDA is performed on the training set based on the speaker labels to obtain a speaker-discriminative feature transformation. All the augmented feature vectors are then mapped by this feature transformation to a low-dimensional speaker-discriminative feature space. The dimensionality of the low-dimensional feature space is 40 in our experiments. Note that this is a significant place where semi-supervision comes in.

The entire training set is used to train a UBM via the EM algorithm, and a GMM mean supervector is obtained for every utterance in the training set via MAP adaptation. Based on the training set, an eigenvoice space and a fishervoice space are learned by PCA and LDA, respectively.

A test set is selected from the GALE Mandarin dataset in a way that there are no overlapping speakers between the test set and the training set. The test set contains 603 speakers and 19023 utterances. On average, each speaker in the test set contains about 30 utterances, and each utterance is about 3-4 seconds long. A GMM mean supervector is obtained for every utterance in the test set in the same way as is for the training set. The GMM mean supervectors of the test utterances are projected onto the eigenvoice space and fishervoice space, respectively. The hierarchical agglomerative (bottom-up) clustering technique is used to perform speaker clustering in the respective subspaces. In the clustering algorithm, the Euclidean distance metric and the Ward's linkage method are used.

Our experiments are carried out in the following manner.

- A case is defined to be an experiment associated with a specific number of test speakers, which are 2, 5, 10, 20, 50, and 100, respectively.
- For each case, this number of speakers are drawn randomly

Table 1. *Average of speaker clustering accuracies over 100 trials for various cases with specific numbers of test speakers (%).*

# test speakers	2	5	10	20	50	100
Original space	96.0	85.0	82.6	78.1	69.4	57.7
Eigenvoice	96.2	85.5	82.9	79.3	69.9	58.5
Fishervoice	**98.4**	**94.0**	**90.8**	**86.6**	**79.6**	**72.3**

from the test set, and all the utterances from the selected speakers are used in the experiment.

- For each case, 100 trial are run, each of which involves a random draw of the test speakers. That is to say, the speakers in different trails are different.
- For each case, the average of the clustering accuracies (the number of correctly clustered utterances over the total number of utterances) over the 100 trials are reported.

The experiment results are shown in Table 1. As we can see from the results, the fishervoice approach significantly outperforms the eigenvoice approach in all cases. The fishervoice approach yields a maximum of 23.6% increase in the average speaker clustering accuracy over the eigenvoice approach in the case of 100 speakers. We also compare the eigenvoice approach to direct clustering in the original speaker space. The results indicate that the eigenvoice approach slightly outperforms direct clustering in the original speaker space, with a maximum of 1.5% increase in the average speaker clustering accuracy in the case of 20 speakers.

6. CONCLUSION

PCA and LDA are two successful linear subspace learning techniques. The speaker subspace learning technique based on PCA is known as the eigenvoice approach. In this paper, we propose a new speaker subspace learning technique based on LDA termed the fishervoice approach. It is shown in our semi-supervised speaker clustering experiments that the fishervoice approach significantly outperforms the eigenvoice approach. This is expected because LDA aims to maximize class separability while seeking a low-dimensional subspace and speaker clustering is essentially a classification problem. It is believed that speaker recognition and diarization would also benefit from the fishervoice approach because both of them are essentially classification problems.

7. REFERENCES

[1] J. P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, 85(9), September 1997, 1437 - 1462.

[2] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," IEEE Trans. ASLP, vol. 14, no. 5, pp. 1557-565, Sept. 2006.

[3] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.L., "Multistage speaker diarization of broadcast news," IEEE Trans. ASLP. vol. 14 no. 5, pp. 1505-1512, Sept. 2006.

[4] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," Lecture Notes in Computer Science, 2007.

[5] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," Proc. DARPA Speech Recognition Workshop'97.

[6] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," Proc. ICASSP'98.

[7] D. Reynolds, E. Singer, B. Carson, G. O'Leary, J. McLaughlin, and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," Proc. ICSLP'98.

[8] S. Chen, and P. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," Proc. ICASSP'98.

[9] W. Tsai, S. Cheng, and H. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," Proc. ICSLP'04.

[10] R. Faltlhauser, and G. Ruske, "Robust speaker clustering in eigenspace," Proc. ASRU'01.

[11] I. Lapidot, H. Guterman, and A. Cohen, "Unsupervised speaker recognition based on competition between self organizing maps," IEEE TNN, 13(4):877-887, 2002.

[12] L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[13] H. Hermansky, "Perceptual linear predictive PLP analysis for speech," Journal of the Acoustic Society of America, 87(4):1738–1752, 1990.

[14] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, Issues 1-2, pp. 91-108, August 1995.

[15] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, 2000.

[16] Campbell, W.M., Sturim, D.E., Reynolds, D.A., "Support vector machines using GMM supervectors for speaker verification," Signal Processing Letters 13(5), 308-311, 2006.

[17] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. JohnWiley & Sons, Inc., 2nd edition, 2001.

[18] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[19] I. Joliffe, Principal Component Analysis. Springer-Verlag, 1986.

[20] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE TPAMI, vol. 19, no. 7, July 1997.

[21] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," Proc. ICSLP'98, pp. 1771-1774.

[22] R. Kuhn, J-C Junqua, P. Nguyen, N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," IEEE Trans. On Speech and Audio Processing. Vol.8 (6), pp. 695-706, Nov. 2000.

[23] O. Thyes, R. Kuhn, P. Nguyen, J.-C. Junqua, "Speaker identification and verification using eigenvoices," Proc. ICSLP'00, Vol.2, pp. 242-246.

[24] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," Proc. ICASSP'08.

[25] Stephen M Chu, Hong-Kwang Kuo, Lidia Mangu, Yi Liu, Yong Qin, and Qin Shi, "Recent advances in the IBM GALE mandarin transcription system," Proc. ICASSP'08.

[26] Arthur Dempster, Nan Laird, and Donald Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. of Royal Stat. Society, B, 39(1):1C38, 1977.

[27] A. K. Jain, M.N. Murthy and P.J. Flynn, "Data Clustering: A Review," ACM Computing Reviews, Nov 1999.

[28] Ward, J.H., "Hierarchical grouping to optimize an objective function," JASA 58:236C245, 1963.