Spherical Discriminant Analysis in Semi-supervised Speaker Clustering[*]

Hao Tang
Dept. of ECE
University of Illinois
Urbana, IL 61801, USA
haotang2@ifp.uiuc.edu

Stephen M. Chu
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
schu@us.ibm.com

Thomas S. Huang
Dept. of ECE
University of Illinois
Urbana, IL 61801, USA
huang@ifp.uiuc.edu

Abstract

Semi-supervised speaker clustering refers to the use of our prior knowledge of speakers in general to assist the unsupervised speaker clustering process. In the form of an independent training set, the prior knowledge helps us learn a speaker-discriminative feature transformation, a universal speaker prior model, and a discriminative speaker subspace, or equivalently a speaker-discriminative distance metric. The directional scattering patterns of Gaussian mixture model mean supervectors motivate us to perform discriminant analysis on the unit hypersphere rather than in the Euclidean space, which leads to a novel dimensionality reduction technique called spherical discriminant analysis (SDA). Our experiment results show that in the SDA subspace, speaker clustering yields superior performance than that in other reduced-dimensional subspaces (e.g., PCA and LDA).

1 Introduction

Speaker clustering is a critical part of speaker diarization (a.k.a. speaker segmentation and clustering) (Barras et al., 2006; Tranter and Reynolds, 2006; Wooters and Huijbregts, 2007; Han et al., 2008). Unlike speaker recognition, where we have the training data of a set of known speakers and thus recognition can be done supervised, speaker clustering is usually performed in a completely unsupervised manner. The output of speaker clustering is the internal labels relative to a dataset rather than real speaker identities. An interesting question is: Can we do semi-supervised speaker clustering? That is, can we make use of any available information that can be helpful to speaker clustering?

Our answer to this question is positive. Here, semi-supervision refers to the use of our prior knowledge of speakers in general to assist the unsupervised speaker clustering process. In the form of an independent training set, the prior knowledge helps us learn a speaker-discriminative feature transformation, a universal speaker prior model, and a discriminative speaker subspace, or equivalently a speaker-discriminative distance metric.

2 Semi-supervised Speaker Clustering

A general pipeline of speaker clustering consists of four essential elements, namely feature extraction, utterance representation, distance metric, and clustering. We incorporate our prior knowledge of speakers into the various stages of this pipeline through an independent training set.

2.1 Feature Extraction

The most popular speech features are spectrum-based acoustic features such as mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) coefficients. In order to account for the dynamics of spectrum changes over time, the basic acoustic features are often supplemented by their first and second derivatives. We pursue a different avenue in which we augment the basic acoustic features of every frame with those of the neighboring frames. Specifically, the acoustic features of the current frame and those of the K_L frames

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

Proceedings of NAACL HLT 2009: Short Papers, pages 57–60,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

to the left and K_R frames to the right are concatenated to form a high-dimensional feature vector. In the context-expanded feature vector space, we learn a speaker-discriminative feature transformation by linear discriminant analysis (LDA) based on the known speaker labels of the independent training set. The resulting low-dimensional feature subspace is expected to provide optimal speaker separability.

2.2 Utterance Representation

Deviating from the mainstream "bag of acoustic features" representation where the extracted acoustic features are represented by a statistical model such as a Gaussian mixture model (GMM), we adopt the GMM mean supervector representation which has emerged in the speaker recognition area (Campbell et al., 2006). Such representation is obtained by *maximum a posteriori* (MAP) adapting a universal background model (UBM), which has been finely trained with all the data in the training set, to a particular utterance. The component means of the adapted GMM are stacked to form a column vector conventionally called a GMM mean supervector. In this way, we are allowed to represent an utterance as a point in a high-dimensional space where traditional distance metrics and clustering techniques can be naturally applied. The UBM, which can be deemed as a universal speaker prior model inferred from the independent training set, imposes generic speaker constraints to the GMM mean supervector space.

2.3 Distance Metric

In the GMM mean supervector space, a naturally arising distance metric is the Euclidean distance metric. However, it is observed that the supervectors show strong directional scattering patterns. The directions of the data points seem to be more indicative than their magnitudes. This observation motivates us to favor the cosine distance metric over the Euclidean distance metric for speaker clustering.

Although the cosine distance metric can be used in the GMM mean supervector space, it is optimal only if the data points are uniformly spread in all directions in the entire space. In a high-dimensional space, most often the data lies in or near a low-dimensional manifold or subspace. It is advantageous to learn an optimal distance metric from the data directly.

The general cosine distance between two data points $\mathbf{x}$ and $\mathbf{y}$ can be defined and manipulated as follows.

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= 1 - \frac{\mathbf{x}^T A \mathbf{y}}{\sqrt{\mathbf{x}^T A \mathbf{x}}\sqrt{\mathbf{y}^T A \mathbf{y}}} \qquad (1) \\ &= 1 - \frac{(A^{1/2}\mathbf{x})^T (A^{1/2}\mathbf{y})}{\sqrt{(A^{1/2}\mathbf{x})^T (A^{1/2}\mathbf{x})}\sqrt{(A^{1/2}\mathbf{y})^T (A^{1/2}\mathbf{y})}} \\ &= 1 - \frac{(W^T\mathbf{x})^T (W^T\mathbf{y})}{\sqrt{(W^T\mathbf{x})^T (W^T\mathbf{x})}\sqrt{(W^T\mathbf{y})^T (W^T\mathbf{y})}} \end{aligned}$$

The general cosine distance can be casted as the cosine distance between two transformed data points $W^T\mathbf{x}$ and $W^T\mathbf{y}$ where $W^T = A^{1/2}$. In this sense, learning an optimal distance metric is equivalent to learning an optimal linear subspace of the original high-dimensional space.

3 Spherical Discriminant Analysis

Most existing linear subspace learning techniques (e.g. PCA and LDA) are based on the Euclidean distance metric. In the GMM mean supervector space, we seek to perform discriminant analysis in the cosine distance metric space. We coin the phrase "spherical discriminant analysis" to denote discriminant analysis on the unit hypersphere. We define a projection from a d-dimensional hypersphere to a d'-dimensional hypersphere where $d' < d$

$$\mathbf{y} = \frac{W^T\mathbf{x}}{\|W^T\mathbf{x}\|} \qquad (2)$$

We note that such a projection is nonlinear. However, under two mild conditions, this projection can be linearized. One is that the objective function for learning the projection only involves the cosine distance. The other is that only the cosine distance is used in the projected space. In this case, the norm of the projected vector $\mathbf{y}$ has no impact on the objective function and distance computation in the projected space. Thus, the denominator term of Equation 2 can be safely dropped, leading to a linear projection.

3.1 Formulation

The goal of SDA is to seek a linear transformation W such that the average within-class cosine similarity of the projected data set is maximized while the average between-class cosine similarity of the projected data set is minimized. Assuming that there are c classes, the average within-class cosine similarity can be written in terms of the unknown projection matrix W and the original data points $\mathbf{x}$

$$S_W = \frac{1}{c}\sum_{i=1}^{c} S_i \tag{3}$$

$$\begin{aligned} S_i &= \frac{1}{|D_i||D_i|} \sum_{\mathbf{y}_j,\mathbf{y}_k \in D_i} \frac{\mathbf{y}_j^T \mathbf{y}_k}{\sqrt{\mathbf{y}_j^T \mathbf{y}_j}\sqrt{\mathbf{y}_k^T \mathbf{y}_k}} \\ &= \frac{1}{|D_i||D_i|} \sum_{\mathbf{x}_j,\mathbf{x}_k \in D_i} \frac{\mathbf{x}_j^T W W^T \mathbf{x}_k}{\sqrt{\mathbf{x}_j^T W W^T \mathbf{x}_j}\sqrt{\mathbf{x}_k^T W W^T \mathbf{x}_k}} \end{aligned}$$

where $|D_i|$ denotes the number of data points in the i^{th} class. Similarly, the average between-class cosine similarity can be written in terms of W and $\mathbf{x}$

$$S_B = \frac{1}{c(c-1)}\sum_{m=1}^{c}\sum_{n=1}^{c} S_{mn} \quad (m \neq n) \tag{4}$$

$$\begin{aligned} S_{mn} &= \frac{1}{|D_m||D_n|} \sum_{\substack{\mathbf{y}_j \in D_m \\ \mathbf{y}_k \in D_n}} \frac{\mathbf{y}_j^T \mathbf{y}_k}{\sqrt{\mathbf{y}_j^T \mathbf{y}_j}\sqrt{\mathbf{y}_k^T \mathbf{y}_k}} \\ &= \frac{1}{|D_m||D_n|} \sum_{\substack{\mathbf{x}_j \in D_m \\ \mathbf{x}_k \in D_n}} \frac{\mathbf{x}_j^T W W^T \mathbf{x}_k}{\sqrt{\mathbf{x}_j^T W W^T \mathbf{x}_j}\sqrt{\mathbf{x}_k^T W W^T \mathbf{x}_k}} \end{aligned}$$

where $|D_m|$ and $|D_n|$ denote the number of data points in the m^{th} and n^{th} classes, respectively.

The SDA criterion is to maximize S_W while minimizing S_B

$$W = arg \max_W (S_W - S_B) \tag{5}$$

Our SDA formulation is similar to the work of Ma et al. (2007). However, we solve it efficiently in a general dimensionality reduction framework known as graph embedding (Yan et al., 2007).

3.2 Graph Embedding Solution

In graph embedding, a weighted graph with vertex set X and similarity matrix S is used to characterize certain statistical or geometrical properties of a data set. A vertex in X represents a data point and an entry s_{ij} in S represents the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$. For a specific dimensionality reduction algorithm, there may exist two graphs. The intrinsic graph $\{X, S^{(i)}\}$ characterizes the data properties that the algorithm aims to preserve and the penalty graph $\{X, S^{(p)}\}$ characterizes the data properties that the algorithm aims to avoid. The goal of graph embedding is to represent each vertex in X as a low dimensional vector that preserves the similarities in S. The objective function is

$$W{=}arg \min\nolimits_W \textstyle\sum_{i \neq j} \|f(\mathbf{x}_i,W){-}f(\mathbf{x}_j,W)\|^2 (s_{ij}^{(i)}{-}s_{ij}^{(p)}) \tag{6}$$

where $f(\mathbf{x}, W)$ is a general projection with parameters W. If we take the projection to be of the form in Equation 2, the objective function becomes

$$W{=}arg \min\nolimits_W \textstyle\sum_{i \neq j} \left\| \frac{W^T \mathbf{x}_i}{\|W^T \mathbf{x}_i\|} - \frac{W^T \mathbf{x}_j}{\|W^T \mathbf{x}_j\|} \right\|^2 (s_{ij}^{(i)}{-}s_{ij}^{(p)}) \tag{7}$$

It is shown that the solution to the graph embedding problem of Equation 7 may be obtained by a steepest descent algorithm (Fu et al., 2008). If we expand the L_2 norm terms of Equation 7, it is straightforward to show that Equation 7 is equivalent to Equation 5 provided that the graph weights are set to proper values, as follows.

$$\begin{aligned} s_{jk}^{(i)} &\leftarrow \frac{1}{c|D_i||D_i|} \quad \text{if } \mathbf{x}_j, \mathbf{x}_k \in D_i, \quad i = 1, ..., c \\ s_{jk}^{(p)} &\leftarrow \frac{1}{c(c-1)|D_m||D_n|} \quad \text{if } \mathbf{x}_j \in D_m, \mathbf{x}_k \in D_n \\ & \quad m, n = 1, ..., c, m \neq n \end{aligned} \tag{8}$$

That is, by assigning appropriate values to the weights of the intrinsic and penalty graphs, the SDA optimization problem in Equation 5 can be solved within the elegant graph embedding framework.

4 Experiments

Our speaker clustering experiments are based on a test set of 630 speakers and 19024 utterances selected from the GALE database (Chu et al., 2008), which contains about 1900 hours of broadcasting news speech data collected from various TV programs. An independent training set of 498 speakers and 18327 utterances is also selected from the GALE database. In either data set, there are an average of 30-40 utterances per speaker and the average duration of the utterances is about 3-4 seconds. Note that there are no overlapping speakers in the two data sets – speakers in the test set are not present in the independent training set.

The acoustic features are 13 basic PLP features with cepstrum mean subtraction. In computing the LDA feature transformation using the independent training set, K_L and K_R are both set to 4, and the dimensionality of the low-dimensional feature space is set to 40. The entire independent training set is used to train a UBM via the EM algorithm, and a GMM mean supervector is obtained for every utterance in the test set via MAP adaptation. The trained UBM has 64 mixture components. Thus, the dimension of the GMM mean supervectors is 2560.

We employ the hierarchical agglomerative clustering technique with the "ward" linkage method. Our experiments are carried out as follows. In each experiment, we perform 4 cases, each of which is associated with a specific number of test speakers, i.e., 5, 10, 20, and 50, respectively. In each case, the corresponding number of speakers are drawn randomly from the test set, and all the utterances from the selected speakers are used for clustering. For each case, 100 trials are run, each of which involves a random draw of the test speakers, and the average of the clustering accuracies across the 100 trials is recorded.

First, we perform speaker clustering in the original GMM mean supervector space using the Euclidean distance metric and the cosine distance metric, respectively. The results indicate that the cosine distance metric consistently outperforms the Euclidean distance metric. Next, we perform speaker clustering in the reduced-dimensional subspaces using the eigenvoice (PCA) and fishervoice (LDA) approaches, respectively. The results show that the fishervoice approach significantly outperforms the eigenvoice approach in all cases. Finally, we perform speaker clustering in the SDA subspace. The results demonstrate that in the SDA subspace, speaker clustering yields superior performance than that in other reduced-dimensional subspaces (e.g., PCA and LDA). Table 1 presents these results.

Metric	Subspace	5	10	20	50
Euc	Orig	85.0	82.6	78.1	69.4
	PCA	85.5	82.9	79.3	69.9
	LDA	94.0	90.8	86.6	79.6
Cos	Orig	90.7	86.5	82.2	77.7
	SDA	**98.0**	**94.7**	**90.0**	**85.9**

Table 1: Average speaker clustering accuracies (unit:%).

5 Conclusion

This paper proposes semi-supervised speaker clustering in which we learn a speaker-discriminative feature transformation, a universal speaker prior model, and a speaker-discriminative distance metric through an independent training set. Motivated by the directional scattering patterns of the GMM mean supervectors, we peroform discriminant analysis on the unit hypersphere rather than in the Euclidean space, leading to a novel dimensionality reduction technique "SDA". Our experiment results indicate that in the SDA subspace, speaker clustering yields superior performance than that in other reduced-dimensional subspaces (e.g., PCA and LDA).

References

C. Barras, X. Zhu, S. Meignier, and J. Gauvain. 2006. Multistage speaker diarization of broadcast news. *IEEE Trans. ASLP*, 14(5):1505–1512.

W. Campbell, D. Sturim, D. Reynolds. 2006. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters* 13(5):308-311.

S. Chu, H. Kuo, L. Mangu, Y. Liu, Y. Qin, and Q. Shi. 2008. Recent advances in the IBM GALE mandarin transcription system. *Proc. ICASSP*.

Y. Fu, S. Yan and T. Huang. 2008. Correlation Metric for Generalized Feature Extraction. *IEEE Trans. PAMI* 30(12):2229–2235.

K. Han, S. Kim, and S. Narayanan. 2008. Strategies to Improve the Robustness of Agglomerative Hierarchical Clustering under Data Source Variation for Speaker Diarization. *IEEE Trans. SALP* 16(8):1590–1601.

Y. Ma, S. Lao, E. Takikawa, and M. Kawade. 2007. Discriminant Analysis in Correlation Similarity Measure Space. *Proc. ICML* (227):577–584.

S. Tranter and D. Reynolds. 2006. An Overview of Automatic Speaker Diarization Systems. *IEEE Trans. ASLP*, 14(5):1557–1565.

C. Wooters and M. Huijbregts. 2007. The ICSI RT07s Speaker Diarization System. *LNCS*.

S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin. 2007. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans. PAMI* 29(1):40–51.