Unsupervised Speaker Diarization Using Riemannian Manifold Clustering

Che-Wei Huang, Bo Xiao, Panayiotis G. Georgiou, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL),
University of Southern California, Los Angeles, CA
{cheweihu, boxiao}@usc.edu, {georgiou, shri}@sipi.usc.edu

Abstract

We address the problem of speaker clustering for robust unsupervised speaker diarization. We model each speaker-homogeneous segment as one single full multivariate Gaussian probability density function (pdf) and take into consideration the Riemannian property of Gaussian pdfs. By assuming that segments from different speakers lie on different (possibly intersected) sub-manifolds of the manifold of Gaussian pdfs, we formulate the original problem as a Riemannian manifold clustering problem. To apply the computationally simple Riemannian locally linear embedding (LLE) algorithm, we impose a constraint on the length of each segment so as to ensure the fitness of single-Gaussian modeling and to increase the chance that all k-nearest neighbors of a pdf are from the same sub-manifold (speaker). Experiments on the microphone-recorded conversational interviews from NIST 2010 speaker recognition evaluation set demonstrate promising results of less than 1% DER.

Index Terms: Speaker Diarization, Riemannian Manifold Clustering, Fisher-Rao metric, Riemannian Locally Linear Embedding (LLE)

1. Introduction

Speaker diarization tries to answer the question of "who spoke when" from audio recordings using speech processing techniques. In general, no prior information about the number of speakers in the conversation or about timing details of turn taking are assumed. Typically, speaker diarization involves two stages: speaker segmentation followed by speaker clustering. Speaker segmentation breaks the entire audio signal into segments that are speaker-homogeneous by performing speaker change detection. Speaker clustering then groups speech segments from the same speaker together while determining the number of speakers present at the same time. These two stages can be performed separately or combined together.

Most state-of-the-art diarization systems are based on a HMM/GMM approach [9, 13]. In such a framework, the original audio signal segmented by Bayesian information criterion (BIC) or generalized log-likelihood ratio (GLR) is deemed to have more initial clusters than the true number expected, i.e., each detected segment is assigned a cluster. The second step is agglomerative hierarchical clustering (AHC) that merges two similar clusters in each iteration until some stopping criterion is reached. Each cluster is subsequently modeled with a mixture of Gaussian pdfs (GMM), and when merged, a new GMM is re-estimated based on the merged data. Viterbi re-segmentation is performed subsequently to refine the segment boundaries using the new models.

Recently, an information theoretic approach based on rate distortion theory [14] has been proposed and proved to be comparable to the state-of-the-art systems with significant savings in computation compared to the HMM/GMM approaches. An alternative based on Variational Bayes is presented in [11, 15, 16] to incorporate the notion of eigenvoices in diarizing telephone conversations. Yet another approach [17] has proposed spectral clustering of segments based on the Kullback-Leibler divergence between the GMM models built on the segments. All of the above methods require an Expectation-Maximization (EM) algorithm at some point in the modeling.

Inspired by the spectral clustering approach, we present a novel geometric algorithm for speaker diarization based on the unsupervised Riemannian locally linear embedding (LLE) framework developed in [7, 8], which has demonstrated impressive performance in computer vision tasks. Almost all existing manifold learning algorithms are neighborhood based and sensitive to the neighborhood, including the Riemannian LLE. However, we will demonstrate that with a simple constraint on the length of segments, we can reduce the sensitivity at the cost of slight increase in computation, and make the algorithm EM-free. In order to focus on speaker clustering, in our proposed algorithm in this paper, we assume the number of speakers is known and each speech segment is speaker-homogeneous.

We highlight our contributions in this paper in the next section, and give a basic review of the mathematical foundation for manifold learning in section 3. In section 4, we present our diarization algorithm. In Section 5 and 6 we present the experiments on the NIST 2010 microphone-recorded conversational interview data set and a discussion.

2. Contributions

Our novel contributions in this work are two-fold: First, to the best of our knowledge, this is the first work to apply manifold clustering with an exact Riemannian metric to speaker diarization, despite the multitude of existing methods in this area. Second, we provide a simple but efficient solution: (i) the proposed solution can robustly work when there is insufficient data to represent local geometry on a statistical manifold; (ii) it also avoids the drawbacks along with the Expectation-Maximization algorithm.

3. Mathematical Formulation for Manifold Learning

3.1. Riemannian Geometry

This section gives a review of the basic terminologies in Riemannian geometry as the mathematical foundation for our diarization algorithm. For more detailed explanation, please refer to [2].

Intuitively, a differentiable manifold $\mathcal{M}$ of dimension d is a space that "locally" resembles a copy of the Euclidean space of the same dimension but not necessarily globally. The concept of the tangent space manifests the locally Euclidean property. Suppose $\mathbf{x} \in M$ and suppose $\alpha(t) : [-\epsilon, \epsilon] \to \mathcal{M}$ is a curve on $\mathcal{M}$ for some $\epsilon > 0$ with $\alpha(0) = \mathbf{x}$. The tangent space at $\mathbf{x}$ on $\mathcal{M}$, denoted by $T_{\mathbf{x}}\mathcal{M}$, is a vector space of tangent vectors $\dot{\alpha}(0)$ for all curves $\alpha(t)$ that pass through $\mathbf{x}$. A tangent space $T_{\mathbf{x}}\mathcal{M}$ can be viewed as a local linearization of the manifold around $\mathbf{x}$. A Riemannian metric on a differentiable manifold

$\mathcal{M}$ associates to each $\mathbf{x} \in \mathcal{M}$ a differentiable inner product $\langle \cdot, \cdot \rangle_{\mathbf{x}}$ on $T_{\mathbf{x}}\mathcal{M}$ and the induced norm by a Riemannian metric is defined as $\|\mathbf{v}\|_{\mathbf{x}}^2 = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{x}}$ $\forall \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. A manifold with a Riemannian metric is called a Riemannian manifold. Define the minimum distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ on the manifold $\mathcal{M}$ to be the Riemannian distance and denote it by dist($\mathbf{x}_i$,$\mathbf{x}_j$). Among all smooth curves from $\mathbf{x}_i$ to $\mathbf{x}_j$, the one with minimum Riemannian distance is called the geodesic curve from $\mathbf{x}_i$ to $\mathbf{x}_j$. For any $\mathbf{x} \in \mathcal{M}$ and any $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$, there exists $\epsilon > 0$ and a geodesic curve $\gamma_{\mathbf{v}}(t)$ for all $|t| < \epsilon$ such that $\gamma_{\mathbf{v}}(0) = \mathbf{x}$ and $\dot{\gamma}_{\mathbf{v}}(0) = \mathbf{v}$.

Now consider the exponential map $\exp_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \to \mathcal{M}$ sending a tangent vector $\mathbf{v}$ to $\gamma_{\mathbf{v}}(1)$, $\exp_{\mathbf{x}}(\mathbf{v}) = \gamma_{\mathbf{v}}(1)$. The inverse of $\exp_{\mathbf{x}}$ is the logarithm map $\log_{\mathbf{x}} : \mathcal{M} \to T_{\mathbf{x}}\mathcal{M}$. For any two points $\mathbf{x}_i$, $\mathbf{x}_j \in \mathcal{M}$, the tangent vector to the geodesic curve from $\mathbf{x}_i$ to $\mathbf{x}_j$ is given by $\mathbf{v} = \log_{\mathbf{x}_i}(\mathbf{x}_j)$ and $\exp_{\mathbf{x}_i}(\log_{\mathbf{x}_i}(\mathbf{x}_j)) = \mathbf{x}_j$. The Riemannian distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = \| \log_{\mathbf{x}_i}(\mathbf{x}_j) \|_{\mathbf{x}_i}$.

3.2. Riemannian Analysis of Probability Density Functions

The application of Riemannian geometry to the manifold of probability density functions plays an important role in this work. Rao [1] first introduced the manifold of statistics where each point on the manifold is a pdf. Geometrization highlights the invariance under coordinate transformation. Rao also showed that the Fisher-Rao metric is a Riemannian metric. In fact, Fisher-Rao metric is the intrinsic unique metric on the statistical manifold and the only metric invariant to re-parametrization [5]. Srivastava et al. [6] gave a "spherical" version of the Fisher-Rao metric that allows closed form expressions for various Riemannian properties. The crux is to re-parametrize pdfs. Let $\mathcal{P}$ be the set of pdfs:

$$\mathcal{P} = \left\{ \mathbf{p} : \mathbb{R} \to \mathbb{R} \middle| \forall s \ \mathbf{p}(s) \geq 0, \int_{\mathbb{R}} \mathbf{p}(s) ds = 1 \right\}.$$

For $\mathbf{v}_1, \mathbf{v}_2 \in T_{\mathbf{p}}\mathcal{P}$, the Fisher-Rao metric [1] is defined as:

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle_{\mathbf{p}} = \int_{\mathbb{R}} \mathbf{v}_1(s) \mathbf{v}_2(s) \frac{1}{\mathbf{p}(s)} ds.$$

It turns out this formulation is hard to deal with because it is not easy to maintain the geodesic curve of any two points of $\mathcal{P}$ to lie on $\mathcal{P}$ under this Fisher-Rao metric [6].

Consider another representation for the set of pdfs.

$$\mathbf{\Psi} = \left\{ \psi : \mathbb{R} \to \mathbb{R} \middle| \forall s \ \psi(s) \geq 0, \int_{\mathbb{R}} \psi^2(s) ds = 1 \right\},$$

is called the *square-root re-parametrization* of pdfs. Rather than a pdf, a point on the statistical manifold now represents a positive square-root of a pdf. It can be shown that the Fisher-Rao metric for the square-root re-parametrization becomes [18]:

$$\langle \mathbf{w}_1, \mathbf{w}_2 \rangle_{\psi} = \int_{\mathbb{R}} \mathbf{w}_1(s) \mathbf{w}_2(s) ds,$$

where $\psi \in \mathbf{\Psi}$ and $\mathbf{w}_1, \mathbf{w}_2 \in T_{\psi}\mathbf{\Psi}$. With such re-parametrization, the Fisher-Rao metric is equivalent to $\mathbb{L}^2$ metric and the set of pdfs is a unit sphere in a Hilbert space. The differential-geometrical properties of a sphere immediately apply. For any two points ψ_1, ψ_2 on a unit sphere, the Riemannian distance dist(ψ_1, ψ_2) is the angle between them:

$$\text{dist}(\psi_1, \psi_2) = \cos^{-1} \langle \psi_1, \psi_2 \rangle$$

The geodesic curve between ψ_1 and ψ_2 is

$$\gamma(t) = \frac{(1-t)\psi_1 + t\psi_2}{t^2 + (1-t)^2 + 2t(1-t)\langle \psi_1, \psi_2 \rangle},$$

and the exponential and the logarithm maps are

$$\begin{aligned} \exp_{\psi_1}(\mathbf{w}) &= \cos(\|\mathbf{w}\|_{\psi_1})\psi_1 + \sin(\|\mathbf{w}\|_{\psi_1}) \frac{\mathbf{w}}{\|\mathbf{w}\|_{\psi_1}} \\ \log_{\psi_1}(\psi_2) &= \frac{\mathbf{u}}{\left(\int_{\mathbb{R}} \mathbf{u}^2(s) ds\right)^{1/2}} \cos^{-1} \langle \psi_1, \psi_2 \rangle \end{aligned}$$

where

$$\langle \psi_1, \psi_2 \rangle = \int_{\mathbb{R}} \psi_1(s) \psi_2(s) ds,$$
$$\mathbf{u} = \psi_2 - \langle \psi_1, \psi_2 \rangle \psi_1, \quad \mathbf{w} \in T_{\psi_1}\mathbf{\Psi}.$$

3.3. Euclidean and Riemannian LLE

With the appropriate metric introduced, this section discusses the clustering framework: one of the best known nonlinear dimensionality reduction techniques, locally linear embedding (LLE), and its generalization to accomodate Riemannian manifolds [7, 8].

Nonlinear dimensionality reduction techniques can be broadly divided into two categories, namely global and local. ISOMAP [4] and LLE [3] are arguably the representatives for these respective categories. The difference between global and local techniques is the scope of properties they intend to preserve.

LLE tries to uncover the nonlinear local embedding of high dimensional data while preserving the neighborhood. The philosophy is to compute similarities between points within a predefined neighborhood size and to apply the spectral clustering on the learned similarity matrix. In general, the free parameters are the size of neighborhood k and the intrinsic dimension d. Since the focus in this work is to apply the LLE algorithm to clustering of speaker segments, the neighborhood size k is the only parameter. Following is the outline of the LLE algorithm for clustering:

1. For each data point $\mathbf{x}_i, i = 1, \cdots, n$, find its k-nearest neighbors based on Euclidean distance. Denote N(i) the index set of k-nearest neighbors of $\mathbf{x}_i$.
2. Construct the similarity matrix W by solving the minimization problem for each $\mathbf{x}_i$

$$\min_{\mathbf{w}_i} \| \sum_{j=1}^{n} w_{ij} \mathbf{x}_j - \mathbf{x}_i \|^2 \quad \text{s.t.} \quad \sum_{j=1}^{n} w_{ij} = 1,$$

   where $w_{ij} = 0$ if $j \notin \text{N}(i)$. $\mathbf{w}_i$ is the i-the row of W.
3. Solve the eigen-decomposition problem of matrix $M \in \mathbb{R}^{n \times n}$, where $M = (I - W)^T (I - W)$, $I \in \mathbb{R}^{n \times n}$ the identity matrix. Select eigenvectors corresponding to the second to the $(m+1)$-th smallest eigenvalues $Z = [\mathbf{u}_1, \cdots, \mathbf{u}_m]$ and apply k-means algorithm to assign rows of Z into m clusters. If i-th row is assigned to cluster m_i where $m_i \in [1, m]$, then $\mathbf{x}_i$ belongs to cluster m_i as well.

The second step is simply a constrained 2-norm minimization, which can be easily solved by Lagrangian multiplier method. The solution is

$$[w_{ii_1}, \cdots, w_{ii_l}, \cdots, w_{ii_k}] = \frac{\mathbf{1}^T C_i^{-1}}{\mathbf{1}^T C_i^{-1} \mathbf{1}}, \quad \forall i_l \in \text{N}(i), \quad (1)$$

where $\mathbf{1}$ is the vector of all ones and C_i the local Gram matrix at $\mathbf{x}_i$ defined as $C_i(j,l) = (\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_l)$.

To generalize the LLE algorithm to incorporate Riemannian structure [7, 8], some modifications to the minimization problem are necessary. First, replacement of the Euclidean distance by the Riemannian distance accounts for the nonlinearity. Next, follow the same spirit to compute the similarities for points on

the statistical manifold. The geodesic linear interpolation of $\mathbf{x}_i$ by all data points in the neighborhood can be expressed as

$$\hat{\mathbf{x}}_i = \exp_{\mathbf{x}_i}\left(\sum_{j=1}^{n} w_{ij}\log_{\mathbf{x}_i}(\mathbf{x}_j)\right).$$

Then solve the following minimization of Riemannian distance square for $\mathbf{w}_i$:

$$\min_{\mathbf{w}_i}\left\|\log_{\mathbf{x}_i}\left(\exp_{\mathbf{x}_i}\left(\sum_{j=1}^{n} w_{ij}\log_{\mathbf{x}_i}(\mathbf{x}_j)\right)\right)\right\|_{\mathbf{x}_i}^2$$

subject to the same constraint on $\mathbf{w}_i$ as in the Euclidean case $\forall 1 \le i \le n$. Since $\exp_{\mathbf{x}_i}$ and $\log_{\mathbf{x}_i}$ are inverse to each other, the objective function can be simplified to

$$\min_{\mathbf{x}_i}\|\sum_{j=1}^{n} w_{ij}\log_{\mathbf{x}_i}(\mathbf{x}_j)\|_{\mathbf{x}_i}^2.$$

The solution by the Lagrangian multiplier method gives the same formula as in Eq.(1) with the local Gram matrix replaced by $C_i(j,l) = \langle\log_{\mathbf{x}_i}(\mathbf{x}_j), \log_{\mathbf{x}_i}(\mathbf{x}_l)\rangle_{\mathbf{x}_i}$.

4. Riemannian Manifold Based Diarization

4.1. Two Challenges

Some observations and issues deserve attention here. First of all, due to the existence of the closed-form expression, the algorithm is computationally simple. There are no iterations involved and certainly there are no convergence issues. However, in order to apply LLE, the choice of neighborhood size is critical to the accuracy of clustering [19]. If the size is too small, the similarity matrix loses the ability to capture sufficient information on the local geometry, especially when the neighborhood size is smaller than the intrinsic dimension. On the other hand, if the size is too large, there is a risk of including points from other sub-manifolds into the same neighborhood on which a similarity matrix is falsely computed. It is also possible [19] that for every neighborhood size, there is always a mixture of points from different sub-manifolds. In such case, the insufficiency of sample points has a great impact on the accuracy.

Challenge 1: As we have discussed above, the neighborhood size is critical to accuracy. However, it can happen that for each sub-manifold, there is only a small number of sample points but the intrinsic sub-manifolds are of high-dimension. A large neighborhood size is thus necessary to capture well the local geometry. The difficulty comes from the fact that due to the small number of sample points, a large neighborhood size tends to include points from different sub-manifolds, which will result in a poorly learned similarity matrix. This is an unwanted situation.

Challenge 2: The second challenge is on the computational aspect. Usually, speech signals are modeled by GMMs, instead of a single Gaussian pdf, to account for the diversity of human speech production. The most common GMM training tool is the well-known EM algorithm, which has certain disadvantages like local maxima and slow convergence. Therefore, it is undesirable to integrate the EM algorithm into the Riemannian LLE.

4.2. Length Constraint

Fortunately, we can address both of these challenges by imposing a length constraint on each segment. Although it increases the size of problem by a factor depending on the turn taking frequency, the slight increase in computation indeed pays off. For small segments as short as one second, a single Gaussian is sufficient to capture the variability of the signal. Thus the issues of local maxima and iterative computations disappear. Second, since points from the same speaker still lie close to each other because they are generated from the same speaker, each speaker now is represented by a denser set of samples. Therefore a large neighborhood size now incurs a smaller risk of connecting points from different sub-manifolds, and the learned similarity matrix based on this larger neighborhood can better capture the local geometry. Experimentally, we set the length constraint to be one second.

4.3. Neighborhood Size k

The choice of k is generally heuristic and data dependent. Elhamifar et al. [19] and Cetingul et al. [20] proposed an automatic way to select k for each sample point $\mathbf{x}_i$ by adding a $l1$-norm regularization to the minimization problem in the formulation of the LLE. The original constrained 2-norm minimization becomes a constrained least squared $l1$ regularization problem (constrained LASSO), which can be processed as quadratic programming and solved by algorithms such as interior-point method. However, importing sparsity constraint into the LLE problem leads to an overhead that at each point the local Gram matrix is of $\mathbb{R}^{N^2}$ rather than of $\mathbb{R}^{k^2}$, and the amount of computation time becomes prohibitive. Moreover, the time spent in solving the convex minimization problem turns out to be a severe bottleneck for the entire clustering problem. Although we highly prefer less supervision and less tuning, for now we do not adopt this sparse approach. While k is a tunable parameter, we will show that for a wide range of k, the performance remains stable below 1%.

4.4. Algorithms

4.4.1. Algorithm 1

Directly apply the Riemannian LLE [7, 8] on the speaker clustering problem as the baseline.

1. Input: a sequence of speaker-homogeneous segments $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, each of which contains a number of acoustic features.
2. For each $\mathbf{x}_i$ model it as a single full multivariate Gaussian pdf of acoustic features and denote the square-root re-parametrization of the pdf by $\mathbf{g}_i$.
3. For each $\mathbf{g}_i$ find the k-nearest neighbors using the Riemannian distance. Denote the index set of k-nearest neighbors of $\mathbf{g}_i$ by N(i).
4. For each $\mathbf{g}_i$ compute the local Gram matrix by $C_i(j,l) = \langle\log_{\mathbf{g}_i}(\mathbf{g}_j), \log_{\mathbf{g}_i}(\mathbf{g}_l)\rangle_{\mathbf{g}_i}$, and solve for $\mathbf{w}_i$ by
$$[w_{ii_1}, \cdots, w_{ii_k}] = \frac{\mathbf{1}^T C_i^{-1}}{\mathbf{1}^T C_i^{-1}\mathbf{1}},$$
$\forall 1 \le l \le k,\ i_l \in$ N(i) and $w_{ij} = 0$ if $j \notin$ N(i), where $\mathbf{1}$ is the vector of all ones.
5. Solve the eigen-decomposition problem for $M = (I - W)^T(I - W)$, where I is the identity in $\mathbb{R}^{N\times N}$ and i-th row of W is $\mathbf{w}_i$. Select the eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_m$ corresponding to the second to the $(m+1)$-th smallest eigenvalues, respectively, and stack them together as $Z = [\mathbf{u}_1, \cdots, \mathbf{u}_m]$, where m is the number of clusters.
6. Apply k-means clustering to rows of Z into m clusters. If i-th row is assigned to cluster j, then $\mathbf{g}_i$ belongs to cluster j, and therefore, $\mathbf{x}_i$ lies on the sub-manifold j. Assign $\mathbf{x}_i$ to cluster j.
7. Output: assignments of segments $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.

4.4.2. Algorithm 2

Based on Algorithm 1, we present our modified algorithm for speaker diarization.

1. Input: a sequence of speaker-homogeneous segments $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, each of which contains a number of acoustic features.
2. For each speaker-homogeneous segment $\mathbf{x}_i$, slice it into one-second long chunks $\mathbf{x}_{ii_1}, \cdots, \mathbf{x}_{ii_N}$.
3. Repeat step 2 for all $1 \leq i \leq n$. Suppose in total it results in N chunks of one-second long speaker homogeneous segments, where $N = \sum_{i=1}^{n} i_N$. Denote them by $\{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$.
4. Apply Algorithm 1 to the set $\{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$, and take the result as the output assignments.
5. Output: finer segments $\{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ of the original $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ and assignments of $\{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$.

5. Experiments

5.1. Data Preparation and Evaluation Scheme

We used the microphone-recorded conversational interview audio in core condition from the NIST 2010 speaker recognition evaluation plan as our test set. There are two channels denoted by A and B, respectively. Channel A is recorded by a designated room microphone and most of the speech is from the interviewee, while the interviewer in the channel B is recorded by a head-mounted close-talking microphone. We merged these two channels into a mono channel audio stream. The data set contains 2477 audio files of five-minute long, roughly 206 hours in total. The acoustic features in the experiments consisted of 20 dimensional raw MFCC, computed with frame size 40 ms and frame shift 20 ms, without first coefficient and without any type of normalization. The reason for 20 MFCC is because they have been proved to be useful in speaker recognition task [10, 11]. Since the feature dimension was 20, the manifold of Gaussian pdfs had dimensionality of $20+20^2/2+10 = 230$(the mean and the upper triangle of the covariance). In our experiments, we took advantage of the speaker segmentation information within the data set such that each segment was speaker-homogeneous. We thus concentrated on the speaker clustering part in focus of our contribution in this paper. Our diarization performance evaluation followed the two-speaker conversation method employed in literature, e.g. [11, 12]. Although this method makes decision on overlapped segments when clustering, they are ignored during evaluation. Because our algorithm takes one segment as a unit, we did not require any "no-score zones" around the segment boundaries. The diarization error rate (DER) in this setting boils down to mis-clustering error rate. The performance varied over the corpus, so we reported the weighted mean and weighted standard deviation of the DER over all files.

5.2. Comparison

For the first experiment, we tested the original Riemannian LLE with no constraint on the length of segments. Each segment was modeled as a single full multivariate Gaussian pdf. Our expectation to the performance was not optimistic due to the limitations of long segments and insufficient samples for the Riemannian LLE, but to our surprise the DER is below 5% from $k = 20$(mean = 4.806%, std = 1.423) to $k = 25$(mean = 4.893%, std = 1.366%). The best is at $k = 23$ with the mean DER 4.607% and the standard deviation 1.283%.

For the result by Algorithm 2, with the constraint on the length of segments, the weighted average error rate is even better, drastically reduced to below 1% for a wide range of k from 35(mean = 0.971%, std = 0.137%) to 75(mean = 0.988%, std = 0.092%). The best performance is found at $k = 57$ with the mean DER 0.881% and the standard deviation 0.093%.

These statistics tell that Algorithm 1 does not learn sufficient information about the local geometry within the samples when k is smaller than 23. As k increases Algorithm 1 gradually gains more accuracy as the similarity matrix improves based on a larger neighborhood size until the effect of connecting points from different sub-manifolds in the neighborhood predominates after k passes 23. However, the optimal k for the algorithm to capture well the local geometry is far away at 57 as seen in Algorithm 2. On the other hand, Algorithm 2 also gains more accuracy as the neighborhood size k increases. Since there are denser samples, the effect of a mixture of samples from different sub-manifolds is none or insignificant for a wide range of k. Hence Algorithm 2 has a better chance to learn the local geometry. This performance of Algorithm 2 supports our proposed scheme by demonstrating the immunity to sparsity in samples and the robustness to the neighboghood size k.

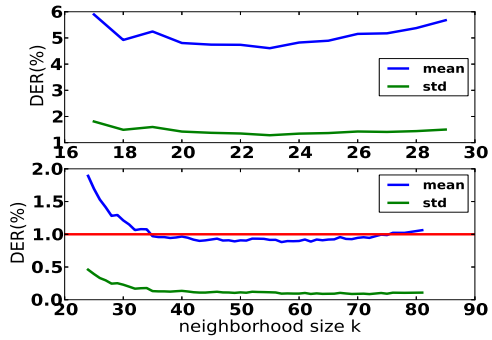

Figure 1: *Performance of Riemannian LLE with (below) and without (above) 1s constraint on the length of segments. The horizontal red line indicates the level of* 1% *DER.*

6. Discussion

The upper plot of Figure 1 justifies the application of the Riemannian LLE to speaker clustering. We attribute the accuracy in Algorithm 1 mostly to the proper Riemannian metric in use. However, the lower plot of Figure 1 shows that Algorithm 2 achieves a much better, more than 80% relative reduction on DER, with only a simple constraint on the segment lengths at a cost of slight increase in computation. We want to note that no post-processing (such as smoothing) has been applied to the output of Algorithm 2. The limitation of the current work is that we have only considered two-speaker conversational interviews, assuming that the the number of speakers and the speaker segmentation information. Such conditions should be relaxed in future research efforts. A method to automatically select the length constraint is important to investigate further.

7. Conclusion

We reported our work on the problem of speaker diarization for the conversational interview corpus. We took a geometric approach, which is distinct from past methodologies, to deal with the problem of speaker clustering. The proposed scheme of constraining length of speech segments served as a simple approach to solve the problem of insufficient local samples and to avoid using the EM algorithm. The proposed diarization algorithm demonstrated the immunity to sparsity in samples and the robustness to the neighboghood size k.

8. ACKNOWLEDGMENT

Special thanks to Mary Francis for her devotion and help in all SAIL research efforts. This research is supported by NSF, NIH and DARPA.

9. References

[1] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters", Bull. Calcutta Math. Soc. 37, 81-89, 1945.

[2] M. P. do Carmo, "Riemannian Geometry", Birkhauser, Boston, 1992

[3] S. T. Roweis and L. K. Saul "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, no. 5500, pp. 2323-2326, Dec 2000.

[4] J. B. Tenenbaum and V. de Silva, J. C. Langford "A global geometric framework for nonlinear dimensionality reduction", Science, vol. 290, no. 5500, pp. 2319-2322, Dec 2000.

[5] N. N. Cencov, "Statistical decision rules and optimal inferences", volume 53 of Translations of Mathematical Monographs, AMS, Providence, USA, 1982.

[6] A. Srivastava, I. Jermyn and S. Joshi "Riemannian analysis of probability density functions with applications in vision", Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, June 2007

[7] A. Goh and R. Vidal, "Clustering and dimensionality reduction on Riemannian manifolds", in: Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, pp. 1-7, 2008.

[8] A. Goh, R. Vidal, "Unsupervised Riemannian clustering of probability density functions", in: W. Daelemans, B. Goethals, K. Morik (Eds.), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, pp. 377392, 2008.

[9] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker diarization: A review of recent research", IEEE Trans. Audio, Speech and Lang. Process., vol. 20, no. 2, pp. 356-370. 2012.

[10] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel "A study of inter-speaker variability in speaker verification", IEEE Trans. Audio, Speech and Lang. Process., vol. 16, no. 5, pp. 980-988, July 2008.

[11] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis", IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 6, pp. 1059-1070, 2010.

[12] "(2001) The NIST year 2001 speaker recognition evaluation plan". [Online] Available: `http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf`

[13] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization", in Interspeech, Lyon, France, pp. 1477-1481, Aug 2013.

[14] D. Vijayasenan, F. Valente, H. Bourlard, "An information theoretic approach to speaker diarization of meeting data", IEEE TASLP, vol. 17, pp. 1382-1393, Sept. 2009

[15] D. Reynolds, P. Kenny, F. Castaldo, "A study of new approaches to speaker dairization", in Proc. Interspeech, ISCA, 2009.

[16] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors", CRIM, Montreal, Technical Report, 2008.

[17] H. Ning, M. Liu, H. Tang, T. Huang, "A spectral clustering approach to speaker diarization", in Proc. Interspeech, ICSLP, Pittsburgh, 2006.

[18] S. Ncube, "A novel Riemannian metric for analyzing spherical functions with applications to HARDI data", Ph.D. disseertation.

[19] E. Elhamifar, R. Vidal, "Sparse manifold clustering and embedding", In NIPS, 2011

[20] H. E. Cetingul, M. J. Wright, P. M. Thompson, R. Vidal, "Segmentation of high angular resolution diffusion MRI using sparse Riemannian manifold clustering", IEEE Trans. Medical Imaging, vol. 33, no. 2, Fed 2014.