Trainable Speaker Diarization

Hagai Aronowitz
IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
haronow@us.ibm.com

Abstract
This paper presents a novel framework for speaker diarization. We explicitly model intra-speaker inter-segment variability using a speaker-labeled training corpus and use this modeling to assess the speaker similarity between speech segments. Modeling is done by embedding segments into a segment-space using kernel-PCA, followed by explicit modeling of speaker variability in the segment-space. Our framework leads to a significant improvement in diarization accuracy. Finally, we present a similar method for bandwidth classification.

Index Terms: speaker segmentation, speaker clustering, speaker diarization, bandwidth classification

1. Introduction
Audio diarization is the process of labeling audio input with labels such as speaker identity, channel type and audio class (speech/music/silence). Audio diarization is a key component for indexing audio archives and for speaker adaptation as part of a transcription system. This paper focuses on broadcast diarization, but the methods developed can be useful for other tasks such as diarization of telephone conversations and meetings diarization.

Speaker diarization is the process of segmenting and labeling audio input according to speakers’ identities. A speaker diarization system usually consists of a speech/non-speech segmentation component, a speaker segmentation component, and a speaker clustering component. In [1] a method for speech/non-speech segmentation based on segmental modeling was introduced. This paper focuses on speaker segmentation and speaker clustering.

Speaker segmentation is the process of identifying change points in an audio input where the identity of the speaker changes. Speaker segmentation is usually done by modeling a speaker with a multivariate normal distribution or with a Gaussian mixture model (GMM) and assuming frame independence. Deciding whether two consecutive segments share the same speaker identity is usually done by applying a Bayesian motivated approach such as Generalized Likelihood Ratio (GLR) [2] or Bayesian Information Criterion (BIC) [3].

Speaker clustering is the process of clustering segments according to speakers’ identity. Speaker clustering is usually based on either the BIC criterion or on Cross Likelihood Ratio (CLR) [4]. A thorough overview of available speaker clustering methods is given in [5].

Lately, anchor modeling, originally introduced for speaker recognition [6], has been successfully used for speaker clustering [7] and for speaker segmentation [8]. Anchor modeling is based on projecting a spoken segment into a space of reference speaker models named anchor-space. A segment is therefore not represented in an absolute way but relatively to a set of speaker models. In [7] a Euclidean distance was used to measure similarity in anchor-space, while in [8] a correlation was used for the same purpose.

In this paper, we suggest to explicitly benefit from available annotated training data to model intra-speaker inter-segment variability. We introduce a novel method for embedding sequences of frames into a space named segment-space and show how to model and classify in segment-space. The embedding is based on kernel-PCA [9]. Given a set of reference speaker models trained on a training corpus, we define a segment-space which is a direct sum of two subspaces. The first subspace named the common-speaker subspace is spanned by the reference speakers. The second subspace named the speaker-unique subspace is the orthogonal complement of the common-speaker subspace, and captures information that does not intersect with the span of the reference speakers. Using kernel-PCA we derive a distance preserving projection from the common-speaker subspace to a Euclidean space, where intra-speaker inter-segment variability can be explicitly modeled. For classification, the information extracted from the common-speaker subspace is integrated with the information extracted from the speaker-unique subspace.

In this paper we present a method of creating a trainable similarity function between segments of speech. This similarity function is a key component in speaker segmentation and speaker clustering algorithms. We evaluate the novel similarity function on a speaker recognition task (where both training and test utterances are 3 sec long) and on a speaker diarization task.

The remainder of this paper is organized as follows: Section 2 introduces the kernel-PCA based approach. In section 3 we describe the experimental setup and results on broadcast news. In section 4 we demonstrate how our techniques can be used for other tasks such as channel detection. Finally, we conclude in section 5.

2. Kernel-PCA based speaker diarization

In this section we introduce a method for embedding sequences of frames into a space named segment-space. The embedding is based on kernel-PCA. We then show how to model speaker variability in segment-space and how to classify in segment-space.

2.1. Kernel-PCA

Kernel-PCA [9] is a kernelized version of the principal component analysis (PCA) algorithm. Function $K(x,y)$ is a kernel if there exists a dot product space $F$ (named ‘feature space’) and a mapping $f : V \rightarrow F$ from observation space $V$ (named ‘input space’) for which:

$$\forall x, y \in V \quad K(x,y) = \langle f(x), f(y) \rangle.$$  \hfill (1)

Given a set of reference vectors $A_1, \ldots, A_n$ in $V$, the kernel-matrix $K$ is defined as $K_{i,j} = K(A_i, A_j)$. The goal of kernel-PCA is to find an orthonormal basis for the subspace spanned by the set of mapped reference vectors $f(A_1), \ldots, f(A_n)$. The
outline of the kernel-PCA algorithm is as follows:
1. Compute a centralized kernel matrix $\tilde{K}$:
   \[
   \tilde{K} = K - 1_n K - K 1_n + 1_n K 1_n
   \]  
   where $1_n$ is an $mn$ matrix with all values set to one.
2. Compute eigenvalues $\lambda_1, \ldots, \lambda_m$ and corresponding eigenvectors $v_1, \ldots, v_m$ for matrix $\tilde{K}$.
3. Normalize each eigenvector by the square root of its corresponding eigenvalue (for the non-zero eigenvalues $\lambda_1, \ldots, \lambda_m$):
   \[
   \tilde{v}_i = v_i / \sqrt{\lambda_i}, \quad i = [1, \ldots, m]
   \]  
   The $i$-th eigenvector in feature space denoted by $f_i$ is:
   \[
   f_i = \langle f(A_1), \ldots, f(A_m) \rangle \tilde{v}_i.
   \]  
   The set of eigenvectors $\{f_1, \ldots, f_m\}$ is an orthonormal basis for the subspace spanned by $\langle f(A_1), \ldots, f(A_m) \rangle$.

Let $x$ be a vector in input space $V$ with a projection in feature space denoted by $f(x)$. $f(x)$ can be uniquely expressed as a linear combination of basis vectors $\{f_i(x)\}$ with coefficients $\{\alpha^x_i\}$, and a vector $u_x$ in $V/\text{span} \{f_1, \ldots, f_m\}$ which is the complementary subspace of span $\{f_1, \ldots, f_m\}$.

\[
 f(x) = \sum_{i=1}^{m} \alpha^x_i f_i + u_x
\]  

Note that $\alpha^x_i = \langle f(x), f_i \rangle$. Using equations (1, 4), $\alpha^x_i$ can be expressed as:

\[
 \alpha^x_i = (K(x, A_1), \ldots, K(x, A_m)) \tilde{v}_i.
\]

We define a projection $T: V \rightarrow \mathbb{R}^m$ as:

\[
 T(x) = (\tilde{v}_1, \ldots, \tilde{v}_m)^T (K(x, A_1), \ldots, K(x, A_m))^T
\]

The following property holds for projection $T$:

if $f(x) = \sum_{i=1}^{m} \alpha^x_i f_i + u_x$ and $f(y) = \sum_{i=1}^{m} \alpha^y_i f_i + u_y$, then:

\[
 \| f(x) - f(y) \|^2 = \| T(x) - T(y) \|^2 + \| u_x - u_y \|^2
\]

Equation (8) implies that projection $T$ preserves distances in the feature subspace spanned by $\{f(A_1), \ldots, f(A_m)\}$.

2.2. Kernel-PCA for speaker diarization

Given a set of sequences of frames corresponding to speaker homogeneous segments, it is desirable to project them into a space where speaker variation can naturally be modeled, while still preserving relevant information. Relevant information is defined in this paper as distances in feature space $F$ defined by a kernel function. Equation (7) suggests such a projection. Using projection $T$ as the chosen projection has the advantage of having $\mathbb{R}^m$ as a natural target space for modeling. Equation (8) quantifies the amount distances are distorted by projection $T$. In order to capture some of the information lost by projection $T$ we define a second projection:

\[
 U(x) = u_x
\]

Although we cannot explicitly apply projection $U$, we can easily calculate the distance between two vectors $u_x$ and $u_y$ using the distance between $x$ and $y$ in feature space $F$ and their distance after projection with $T$:

\[
 \| f(x) - U(y) \|^2 = \| f(x) - f(y) \|^2 - \| T(x) - T(y) \|^2
\]

Using both projections $T$ and $U$ enables capturing the relevant information. The subspace spanned by $\{f(A_1), \ldots, f(A_n)\}$ is named the common-speaker subspace, as attributes that are common to several speakers will typically be projected into it. The complementary space is named the speaker-unique space, as attributes that are unique to a speaker will typically be projected to that subspace.

2.3. Modeling in common-speaker subspace

The purpose of the projection of the common-speaker subspace into $\mathbb{R}^m$ using projection $T$ is to enable modeling of inter-segment speaker variability. Inter-segment speaker variability is closely related to intersession variability modeling which has proven to be extremely successful for speaker recognition [10], [11]. We model speakers’ distributions in common-speaker subspace as multivariate normal distributions with a shared full covariance matrix $\Sigma$ which is $mn$ dimensional ($m$ is the dimension of the common-speaker space).

Given an annotated training dataset, we extract non-overlapping speaker homogeneous segments (of fixed length). Given speakers $s_1, \ldots, s_k$, with $n(s_i)$ segments for speaker $s_i$, $\mathcal{T}(s_{i,1}), \ldots, \mathcal{T}(s_{i,n(s_i)})$ denote the $n(s_i)$ segments of speaker $s_i$ projected into common-speaker subspace. We estimate $\Sigma$ as

\[
 \Sigma = \frac{1}{\sum_{i=1}^{k} n(s_i)} \sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} (\mathcal{T}(s_{i,j}) - \mu_{s_i})(\mathcal{T}(s_{i,j}) - \mu_{s_i})^T
\]

where $\mu_{s_i}$ denotes the mean of the distribution of speaker $s_i$ and is estimated as

\[
 \mu_{s_i} = \frac{1}{n(s_i)} \sum_{j=1}^{n(s_i)} \mathcal{T}(s_{i,j})
\]

We regularize $\Sigma$ by adding a positive noise component $\eta$ to the elements of its diagonal

\[
 \Sigma = \Sigma + \eta I
\]

The resulting covariance matrix is guaranteed to have eigenvalues greater than $\eta$, therefore it is invertible.

Given a pair of segments $x$ and $y$ projected into common-speaker subspace $\mathcal{T}(x)$ and $\mathcal{T}(y)$ respectively, the likelihood of $\mathcal{T}(y)$ conditioned on $\mathcal{T}(x)$ and assuming $x$ and $y$ share the same speaker identity is

\[
 \Pr(\mathcal{T}(y) | \mathcal{T}(x), x - y) = \frac{1}{(2\pi)^{\frac{m^2}{2}}} e^{-\frac{1}{2} \| \mathcal{T}(y) - \mathcal{T}(x) \|^2}
\]
where $2\Sigma$ is the covariance matrix of the random variable $T(y)-T(x)$.

For the sake of efficiency, we diagonalize the covariance matrix $2\Sigma$ by computing its eigenvectors $\{e_i\}$ and eigenvalues $\{\beta\}$. Defining $E$ as $\{e_1^T, \ldots, e_m^T\}$, equation (14) reduces to:

$$
\Pr(T(y)|T(x), x \sim y) = \frac{1}{(2\pi)^{(\frac{1}{2})} \sqrt{\prod_i \beta_i}} \frac{1}{\sqrt{\prod_i \beta_i}} e^{\frac{-1}{2} \sum_{i=1}^{m} (\hat{T}_i(y) - \hat{T}_i(x))^2/\beta_i}.
$$

(15)

where $\hat{T}(x) = E \cdot T(x), \hat{T}(y) = E \cdot T(y)$ and $[x_i]$ is the $i$-th coefficient of $x$.

### 2.4. Modeling in speaker-unique subspace

$\Delta^2_{m}(x,y)$ denotes the squared distance between segments $x$ and $y$ projected into the speaker unique subspace. We assume

$$
\Pr(\Delta^2_{m}(x,y)|x \sim y) = \frac{1}{2\pi \sigma^2} e^{-\frac{1}{2}(\Delta_{m}(x,y))^2 / \sigma^2}
$$

(16)

and estimate $\sigma_m$ from the development data.

### 2.5. Modeling in segment space

The likelihood of segment $y$ given segment $x$ and the assumption that both segments share the same speaker identity is

$$
\Pr(y|x, x \sim y) = \Pr(T(y)|T(x), x \sim y) \Pr(\Delta^2_{m}(x,y)|x \sim y).
$$

(17)

The expression in (17) can be calculated using eqs. (15) and (16).

### 2.6. Score normalization

The speaker similarity score between segments $x$ and $y$ is defined as $\log(Pr(y|x, x \sim y))$. Score normalization is a standard and extremely effective method in speaker recognition. We use T-norm [12] and TZ-norm [10] for score normalization in the context of speaker diarization. Given held out segments $t_1, \ldots, t_f$ from a development set. The T-normalized score $S(y|x)$ of segment $y$ given segment $x$ is:

$$
S(y|x) = \frac{\log(Pr(y|x, x \sim y)) - \text{mean}(\log Pr(y|t_i, t_j \sim y))}{\text{var}(\log Pr(y|t_i, t_j \sim y))}.
$$

(18)

The TZ-normalized score of segment $y$ given segment $x$ is calculated similarly according to [10].

### 2.7. Kernels for speaker diarization

In [13] it was shown that under reasonable assumptions a GMM trained on a test utterance is as appropriate for representing the utterance as the actual test frames (the GMM is approximately a sufficient statistic for the test utterance w.r.t. GMM scoring). Therefore the kernels used are based on GMM parameters trained for the scored segments. GMMs are maximum-posteriori (MAP) adapted from a universal background model (UBM) of order 1024 with diagonal covariance matrices.

The kernel used in this paper was inspired by [14]. The kernel is based on the weighted-normalized GMM means:

$$
K(x, y) = \sum_{g=1}^{G} \sum_{d=1}^{D} \sum_{g'} \frac{\mu_{g,d}^x \mu_{g',d}^y}{\sigma_{g,d}^x \sigma_{g',d}^y}.
$$

(19)

where $\mu_{g,d}^x$ and $\mu_{g,d}^y$ stand for the $d$-th coordinate of the mean of the $g$-th Gaussian of GMMs $x$ and $y$ respectively. $w_{g,UBM}$ and $\sigma_{g,UBM}$ stand for the weight and the $d$-th coordinate of the standard deviation of the $g$-th Gaussian of the UBM.

### 3. Experiments

#### 3.1. Tasks

This paper focuses on the segment scoring component of speaker diarization algorithm and not on the actual segmentation and clustering techniques. Therefore the experiments reported focus on the following task: given a pair of 3sec segments drawn from the same show, the pair should be classified to either ‘same speaker’ or ‘different speaker’. In addition, we report preliminary standard diarization results.

#### 3.2. Anchor modeling based systems

The baseline anchor modeling based system was inspired by [8]. The front-end consists of Mel-frequency cepstrum coefficients (MFCC) with cepstral mean subtraction (CMS). An energy based voice activity detector is used only for CMS. The final feature set is 24 MFCCs + 24 delta MFCCs extracted every 10ms using a 25ms window. Anchor models are trained similarly as described in subsection 2.7. A correlation based distance was used for anchor space scoring as it outperformed slightly using the Euclidean distance.

An additional baseline system was developed using the anchor modeling framework with the following modification: the GMM-log-likelihood ratio based embedding used in [6-8] was replaced by an embedding based on the kernel defined in (19). The following score is used:

$$
S_{\text{kernel}}(x,y) = -K(x,y) + 2K(x,y) - K(y,y).
$$

(20)

#### 3.3. Datasets and Protocol

The approach kernel-PCA based was evaluated on Arabic broadcast news. The GALE Y1Q1-Y1Q4 training datasets [15] were used for training the UBM and for training 700 reference speakers. BNAD05 which is a collection of 12 shows (5.5 hours in total), was used for modeling in segment space and for T-norm and TZ-norm modeling. BNA05 which is a collection of 12 shows (5.5 hours in total) was used for evaluation. The test set contains audio from five different broadcasting networks and has a signal-to-noise ratio which varies significantly from 40db to 10db. Both the BNAD05 and BNA05 were segmented into non-overlapping segments of 3sec. 196 segments were randomly selected (one segment from each speaker in each show) from BNAD05 for T-norm and TZ-norm modeling. The rest of BNAD05 (6272 segments) was used for modeling in segment space. 207 segments were randomly selected (one segment from each speaker in each show) from BNA05 and used as target speaker models. The rest of BNA05 (6756 segments) was used as test segments.
All target speakers models were scored against all test segments from the same show.

3.4. Results

In Table 1 we present the equal error rate (EER) for the baseline anchor modeling system, the anchor modeling system with the kernel based scoring, and the kernel-PCA based systems on a 3sec-3sec speaker recognition task.

<table>
<thead>
<tr>
<th>System</th>
<th>no-norm EER (%)</th>
<th>T-norm EER (%)</th>
<th>TZ-norm EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor modeling</td>
<td>15.1</td>
<td>12.9</td>
<td>17.8</td>
</tr>
<tr>
<td>Anchor modeling + intra-speaker modeling</td>
<td>11.5</td>
<td>10.8</td>
<td>14.9</td>
</tr>
<tr>
<td>Kernel based scoring</td>
<td>16.7</td>
<td>9.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Kernel-PCA projection</td>
<td>16.7</td>
<td>9.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Kernel-PCA projection + intra-speaker modeling</td>
<td>11.8</td>
<td>10.8</td>
<td>14.9</td>
</tr>
</tbody>
</table>

The anchor modeling baseline and the kernel-PCA based system with intra-speaker modeling were independently integrated into a speaker change detection algorithm and an agglomerative clustering algorithm. On BNA05, the kernel-PCA based system achieved a 39% reduction in speaker error rate (SER) [5] compared to the baseline (12.9% → 7.9%).

4. Kernel-PCA based bandwidth detection

The kernel-PCA framework described in section 2 can be applied to any GMM based classification algorithm. The algorithm was applied with minor modifications for bandwidth detection in broadcast news. The definition of the task is as following: given a 3sec segment, the segment should be classified to either ‘narrowband’ (telephone) or ‘wideband’ data. The kernel-PCA based algorithm is based on the segment-space embedding described in subsection 2.1-2.2. The same datasets described in subsection 3.3 were used except for T-norm and TZ-norm which were not applied. In order to exploit the available training data for each channel, several modeling techniques in common-speaker segment space were explored. The results in Table 2 show an improvement over a baseline GMM system using the same setup. The best results were achieved using a GMM of order 2 for modeling each channel. In all experiments, unique-speaker subspace was not modeled.

Table 2. Channel detection results on BNA05 for detection of 3sec segments. Best system is in *Boldface*.

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline GMM</td>
<td>17.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Kernel-PCA projection</td>
<td>18.7</td>
<td>-6.3</td>
</tr>
<tr>
<td>K-PCA + GMM (order=1)</td>
<td>13.8</td>
<td>21.6</td>
</tr>
<tr>
<td>K-PCA + GMM (order=2)</td>
<td>12.2</td>
<td>30.7</td>
</tr>
<tr>
<td>K-PCA + SVM (RBF kernel)</td>
<td>12.9</td>
<td>26.7</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper presented a novel framework for speaker diarization where labeled training data was actively used to explicitly model intra-speaker inter-segment variability. Kernel-PCA was used to embed sequences of frames into a vector space where modeling can be easily done. The kernel-PCA based approach achieved a 51% reduction in EER (15.1→7.4) compared to the anchor modeling based baseline for speaker recognition of pairs of 3sec segments which is a key component in speaker diarization systems. Preliminary speaker diarization experiments using the kernel-PCA based scoring method have shown a 39% reduction in speaker error rate. A similar approach led to a 50.7% reduction in error rate for channel detection and can be used for speaker verification [16], language identification and gender identification.

6. References