T-Test Distance and Clustering Criterion for Speaker Diarization

Trung Hieu Nguyen1,2, Eng Siong Chng2, Haizhou Li1,2

1Department of Human Language Technology, Institute for Infocomm Research, Singapore
2School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

In this paper, we present an application of student’s t-test to measure the similarity between two speaker models. The measure is evaluated by comparing with other distance metrics: the Generalized Likelihood Ratio, the Cross Likelihood Ratio and the Normalized Cross Likelihood Ratio in speaker detection task. We also propose an objective criterion for speaker clustering. The criterion deduces the number of speakers automatically by maximizing the separation between intra-speaker distances and inter-speaker distances. It requires no development data and works well with various distance metrics. We then report the performance of our proposed similarity distance measure and objective criterion in speaker diarization task. The system produces competitive results: low speaker diarization error rate and high accuracy in detecting number of speakers.

Index Terms: speaker diarization, speaker detection, intra-speaker, inter-speaker.

1. Introduction

Speaker diarization is the process to detect speaker turns and to group together segments uttered by the same speaker. Most of the state-of-the-art systems use hierarchical schemes for speaker clustering. Speech segments are split or merged until some desired criteria are satisfied. In doing so, all systems need to define: a distance between clusters/segments and a criterion to define: a distance between clusters/segments and a criterion that indicates when the clusters have the same advantage as BIC as no development data is required.

Many distance measures were proposed in the past and among them, the most popular distances are: the Bayesian Information Criterion (BIC) [1], the Generalized Likelihood Ratio (GLR) [2], the Cross Likelihood Ratio (CLR) [3] and its normalized version the Normalized Cross Likelihood Ratio [4]. Conceptually, all of these measures are operating in the score space with the underlying assumption that if \( \lambda_1 \) and \( \lambda_2 \) are models of the same speaker then the likelihood score value \( L(X|\lambda_1) \) would be close to the likelihood score value \( L(X|\lambda_2) \) where \( X = \{x_1, x_2, \ldots, x_N\} \) are the observed feature vectors. In this paper, we propose a novel distance measure with different assumption: if \( \lambda_1 \) and \( \lambda_2 \) are models of the same speaker then the population of likelihood score values \( \{L(x_i|\lambda_1), \forall x_i \in X\} \) would be close to the population of likelihood score values \( \{L(x_i|\lambda_2), \forall x_i \in X\} \). We expect that this measure would perform better than others because it does not use a single value but captures the statistics of populations. The theory is presented in section 2 and the comparative study is shown in section 5.

In a speaker diarization system, a good clustering criterion is as important as a good distance measure because it decides the final number of speakers. Many systems [5, 6, 7] use clustering criterion with thresholds derived from development set. This approach however presents many issues when there is mismatch between development data and test data. Ajmera [8] proposed a modified version of BIC which can be used as both distance measure and clustering criterion. The algorithm does not require development data, however it is tightly integrated with the BIC distance measure and has the same disadvantages as BIC. Hence, we are motivated to introduce a new clustering criterion based on the maximization of separation between intra-speaker distances and inter-speaker distances. This criterion has the same advantage as BIC as no development data is needed and furthermore the criterion works with many distance metrics. The effectiveness of the criterion only depends on how good the distance metrics are. The theory is presented in section 3 and the experiment is shown in section 5. We then apply our proposed distance measure and clustering criterion in the speaker diarization system. The system description is presented in section 4 and the performance on meeting speech data is reported in section 5.

2. Similarity measure between two speaker models

We first establish some backgrounds and assumptions for our proposed similarity measure: given two probability distribution functions \( f(x) \), \( g(x) \) and a population \( X = \{x_1, x_2, \ldots, x_N\} \).

Denote:

\[
S_f(X) = \{f(x_i)|x_i \in X\} \quad (1)
\]

\[
S_g(X) = \{g(x_i)|x_i \in X\} \quad (2)
\]

We then define the similarity between \( f(x) \), \( g(x) \) as the similarity between two populations \( S_f(X) \) and \( S_g(X) \). We define the similarity between two populations as:

\[
T_d = d(S_f(X), S_g(X)) = \frac{|m_1 - m_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (3)
\]

where \( m_1, \sigma_1, n_1, m_2, \sigma_2, n_2 \) are respectively the mean, standard deviation and size of \( S_f(X) \) and \( S_g(X) \). Smaller value of \( T_d \) indicates \( f(x) \) is more similar to \( g(x) \). This is actually the student’s t-test applied to two populations, however in this case, the normality assumption is not critical for two reasons: the population size of \( X \) is large and we are not interested in the distribution of \( T_d \) but the value of \( T_d \) itself.

Applying the above formulas in the context of measuring the distance between two speakers \( S_1 = \{x_1, x_2, \ldots, x_N\} \) and \( S_2 = \{y_1, y_2, \ldots, y_M\} \), with the following proposed distribution function:

\[
f(x) = \log L(x|\lambda_{C_1}) - \log L(x|\lambda_{UBM}) \quad (4)
\]

\[
g(x) = \log L(x|\lambda_{C_2}) - \log L(x|\lambda_{UBM}) \quad (5)
\]

where \( X = \{x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_M\} \), \( x_i, y_i \) are the feature vectors, \( \lambda_{C_1} \) is the model estimated using feature vectors of speaker \( C_1 \), \( \lambda_{C_2} \) is the model of speaker \( C_2 \), \( \lambda_{UBM} \) is...
is the universal background model and $logL(x|\lambda)$ is the log-likelihood function of feature vector $x$ with model $\lambda$. The distance between speaker $S_1$ and $S_2$ is then computed using (3); a smaller value of $T_d$ indicates that two speakers are more similar to each other. Unlike other distance metrics in which only the difference between the sum or weighted sum of likelihood scores are computed, $T_d$ also takes into account the variance of likelihood scores and the size of clusters. This characterisitcs make $T_d$ more robust and reliable even in the case when there is big mismatch between segment lengths.

3. Criterion for speaker clustering

We begin with the high-level description of the problem: If $X = \{x_1, x_2, \ldots, x_N\}$ is the data where $x_i$ is a feature vector or a set of feature vectors and there are $M$ different ways \{C^{(1)}, C^{(2)}, \ldots, C^{(M)}\} to cluster the data. We want to answer the question which $C^{(i)}$ is the optimal clustering. In this section, we first give the definitions of some terms, then present the criterion for optimal clustering and finally how to use it in our speaker diarization system.

3.1. Intra-cluster distances and inter-cluster distances

Given $C^{(i)}$ is a way to cluster the data $X$ into $K_i$ clusters $C^{(i)} = \{C^{(i)}_1, C^{(i)}_2, \ldots, C^{(i)}_{K_i}\}$. Denote $d(x_m, x_n)$ the distance between $x_m, x_n$ and:

$$D(C_i, C_j) = \{d(x_m, x_n) | x_m \in C_i, x_n \in C_j \forall m, n\}$$

(6)

$$D_{\text{intra}} = \bigcup_{i=1}^{K} D(C_i, C_i)$$

(7)

$$D_{\text{inter}} = \bigcup_{1 \leq i < j \leq K} D(C_i, C_j)$$

(8)

where $D_{\text{intra}}$ is the population of intra cluster distances and $D_{\text{inter}}$ is the population of inter cluster distances.

3.2. Optimal clustering criterion

In this paper, $C^{(i)}$ is considered as the optimal clustering if the separation between two population $D_{\text{intra}}$ and $D_{\text{inter}}$ is maximized. The measure of separation are formulated with two assumptions whether $D_{\text{intra}}, D_{\text{inter}}$ follow Gaussian distributions or not.

3.2.1. With assumption of normality

The distributions of $D_{\text{intra}}$ and $D_{\text{inter}}$ are assumed to be Gaussian. The separation between $D_{\text{intra}}$ and $D_{\text{inter}}$ is defined as:

$$T_s = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(9)

where $m_1, \sigma_1, n_1, m_2, \sigma_2, n_2$ are respectively the mean, standard deviation, size of $D_{\text{intra}}$ and $D_{\text{inter}}$. $T_s$ can be evaluated for any two populations regardless of their distributions, however it is not efficient using $T_s$ to measure the separation if the distributions are very different from normal. The next section will introduce the measure for general cases.

3.2.2. Without assumption of normality

When it is unreasonable to assume normality, we define the separation between two populations as $\rho$ which is a non-parametric measure of the overlap between two distributions. $\rho$ measure is a linear transform of Mann-Whitney U value [9]. The procedures to compute $\rho$ are summarized in the following steps:

- $D = D_{\text{intra}} \cup D_{\text{inter}}$.
- Sort the value of $D$ in ascending order and assign a ranking order for each element of $D$.

$$R_{\text{intra}} = \sum_{x_i \in D_{\text{intra}}} \text{rank}(x_i)$$

(10)

$$U_{\text{intra}} = R_{\text{intra}} - \frac{|D_{\text{intra}}||D_{\text{intra}}|+1}{2}$$

(11)

$$\rho = \frac{|U_{\text{intra}}|}{|D_{\text{intra}}||D_{\text{inter}}| - 0.5} \times 2$$

(12)

where $\text{rank}(x_i)$ is the order of $x_i$ in the sorted sequence of $D$, $||\cdot||$ is the cardinal of the set. $\rho$ can take values between 0 and 1. A $\rho$ of 0 represents complete overlap while a value of 1 represents complete separation of the distributions. $\rho$, unlike $T_s$ is less sensitive to outliers. However, the separation between two populations defined by $\rho$ is not as discriminative as $T_s$ in cases when two sets are completely separated, $\rho$ has value of 1 regardless of the difference between the means of two populations. The performance of these two measurements as our cluster criterion will be reported in several experiments in later sections.

3.3. Speaker clustering in diarization system

In the agglomerative clustering framework, the data is first over-segmented into small clusters and then these small clusters are merged until a stopping criterion is satisfied. In our system, we also first segment the data into small clusters and then execute iterative clustering, merging until there is only one cluster left. For each merging iteration $i$ with the clusters $C^{(i)} = \{C^{(i)}_1, C^{(i)}_2, \ldots, C^{(i)}_{K_i}\}$, we compute the separation between $D_{\text{intra}}$ and $D_{\text{inter}}$ using the following steps:

- Each cluster $C^{(i)}_j$ is divided into segments of $L$ seconds.
- Compute the distance between every pairs of segments using formula (3).
- Acquire $D_{\text{intra}}$ and $D_{\text{inter}}$ as described in (7) and (8).
- Compute the separation between $D_{\text{intra}}$ and $D_{\text{inter}}$ using either $\rho$ or $T_s$.

We then record the separation values at each iteration. The clustering $C^{(i)}$ corresponding to the maximum value of $\rho$ or minimum value of $T_s$ is selected as the final speaker clusters. With this framework, we do not need any development data to detect the number of speakers and the empirical results shown in the later sections confirm that this framework is indeed effective.

4. Speaker Diarization System

The system follows the standard agglomerative clustering framework.

4.1. Preprocessing

Speech data is first processed by applying [10] to obtain both the enhanced speech and speech presence probability of each frame.
4.2. Voice Activity Detection

We extract 19 MFCC features and apply RASTA filter with cepstral-mean normalization (CMN).

Initially, each frame is classified into one of two clusters: speech or non-speech based on the speech presence probability. Then, a model is built for each cluster then Viterbi algorithm is run several times to refine segment boundary and cluster assignment.

4.3. Initialization

After silence removal, the speech data is uniformly divided into segments of 15 seconds and Viterbi resegment is run. We then form clusters of about 30 seconds by merging two segments closest to each other using $T_d$ distance. We repeat this step to obtain initial clusters of approximately 60 seconds each. Due to Viterbi resegment, the initial clusters are generally of different length.

4.4. Iterative clustering

Each cluster is modeled by a GMM model with 32 mixtures, diagonal covariance matrix. At each iteration, the two clusters with smallest $T_s$ are merged and the iterative clustering is executed until there is only one cluster left. The optimal clustering is selected as described in section 3.3.

5. Experiments and Results

5.1. Database

The speaker diarization system was tested on Rich Transcription (RT) 2007 conference data released by NIST for RT07 benchmark on the single distance microphone condition. The database consists of 8 speech audio files from various meeting rooms. A duration of approximate 20 minutes is extracted from each audio file for evaluating. There are total of 35 speakers in the database, and there are between 4 and 6 speakers in each meeting. This database is called set A.

The second database that we work on is NIST Speaker Recognition 2004. We extract 48 speakers in which 34 of them are females from database with the conditions: all conversations are in English, the channel conditions are handset and regular phone. This database is called set B.

5.2. Evaluation criteria

Diarization error rate is a time-based score which calculates the percentage of speaker time which is not mapped correctly to a reference speaker.

$$DER = \frac{\sum_s dur(s)}{\sum_s dur(s)} \cdot \frac{\max(N_{ref}(s), N_{sys}(s)) - N_{correct}(s)}{N_{ref}(s)}$$

where $s$ is the longest continuous segments for which the reference and system speakers do not change, $dur(s)$ is the duration of $s$, $N_{ref}(s)$ is the number of reference speakers in $s$, $N_{sys}(s)$ is the number of system speakers in $s$ and $N_{correct}(s)$ is the number of mapped reference speakers which match the system speakers.

Speaker detection performance is also evaluated in terms of the miss speakers (speaker in reference but not in system) and false alarm (FA) speakers (speaker not in reference but in system).

5.3. Speaker diarization experiments

The system was tested on set A. To detected the number of speakers, we experimented with various segment lengths and show that the algorithm is not too sensitive to the length of segments. The results of two separation measure $\rho$ and $T_s$ is also reported in Table 1.

<table>
<thead>
<tr>
<th>Separation Measure</th>
<th>Segment Length</th>
<th>DER (%)</th>
<th>Miss Speaker</th>
<th>FA Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>10s</td>
<td>20.34</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\rho$</td>
<td>15s</td>
<td>18.75</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$\rho$</td>
<td>20s</td>
<td>22.46</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>$T_s$</td>
<td>10s</td>
<td>19.80</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$T_s$</td>
<td>15s</td>
<td>19.39</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$T_s$</td>
<td>20s</td>
<td>23.39</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

The obtained results are comparable with state-of-the-art systems [11]. Both separation measures perform well for this database and the number of speakers detected is highly accurate. We observed that longer segment length did not perform better, the reason is that each cluster is not guaranteed to contain speech data from only one speaker (cluster impurity), hence there may be many speakers in a long segment and it will affect the overall performance. It is safer to use shorter segments at the cost of higher computation.

5.4. Speaker detection with different distance metrics

This experiment was carried on set B. There are three segments for each speaker with different lengths: 20, 40 and 80 seconds. We measured the distances between every pair of segments and based on the distance values, each pair of segments was judged to be from a same speaker or not. The evaluation criterion is the detection error tradeoff (DET) curve [12] which basically measures the tradeoff between false alarm and miss detection. A good distance measure will produce a DET curve which is close to the origin. We compare our proposed measure $T_d$ with other metrics: GLR, CLR, NCLR and the results are shown in Figure 1. Figure 2 shows the distribution of distances among intra-speakers and inter-speakers.

The proposed measure $T_d$ outperformed other distance metrics as shown in Figure 1. The GLR was significantly worse than the other metrics because it has problem when there is a big mismatch between the length of the segments. Its distributions for same speakers and difference speakers are not well separated. With $T_d$, there is almost no overlapping between two distributions which confirms that $T_d$ is a good distance measure to distinguish among speakers.

5.5. Experiments with clustering criterion

Experiment setup: from the pool of 48 speakers from set B, we selected randomly $K$ speakers ($K$ is between 4 and 8) and form an experiment set. We formed 100 of such experiment sets; the total number of speakers in these 100 experiment sets is 618. The speech data for each speaker consists of several segments of 30s each.

We then perform iterative clustering, the merging decision is based on the reference file to ensure that there is no clustering errors. At each merging iteration, we computed the separation distance $\rho$ and $T_s$. The purpose of these experiments is...
Experimental results are summarized in Table 2. The last column indicates the number of experiment sets out of 100 in which the number of speakers are detected correctly. From Table 2, we observe that $\rho$ and $T_s$ are comparable to each other in terms of number of experiment sets detected correctly, however in terms of miss speakers and false alarm speakers, $T_s$ performs much better than $\rho$. In Figure 1, we note that the optimal speaker detection performance is related to the distance metrics performance. In summary, we have demonstrated that our optimal speaker clustering criterion is justified. The experimental results conform to the theory and the performance of the algorithm is affected by the performance of the distance metrics being used.

6. Conclusion

We have proposed a similarity measure between two speaker models $T_d$ based on student’s t-test. The measure is proved to be very effective in distinguishing among speakers. The performance of $T_d$ in speaker detection task outperforms other distance metrics.

We have also proposed a novel clustering criterion. The criterion is applicable for many common distance metrics and its performance is depending on how good the distance metrics are. When combining with our proposed distance measure $T_d$, the obtained results is very good (99% correct).

The distance measures $T_d$ and clustering criterion are then being used in speaker diarization system. The system has low DER (≈19%) and high accuracy in detecting number of speakers.

7. References


Table 2: Clustering criterion and various distance metrics

<table>
<thead>
<tr>
<th>Separation Measure</th>
<th>Distance Metrics</th>
<th>Missspeaker (%)</th>
<th>FAspeaker (%)</th>
<th>Correct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>GLR</td>
<td>2</td>
<td>388</td>
<td>54</td>
</tr>
<tr>
<td>$\rho$</td>
<td>KL2</td>
<td>1</td>
<td>140</td>
<td>86</td>
</tr>
<tr>
<td>$\rho$</td>
<td>CLR</td>
<td>0</td>
<td>194</td>
<td>85</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$T_d$</td>
<td>0</td>
<td>23</td>
<td>98</td>
</tr>
<tr>
<td>$T_s$</td>
<td>GLR</td>
<td>18</td>
<td>10</td>
<td>79</td>
</tr>
<tr>
<td>$T_s$</td>
<td>KL2</td>
<td>8</td>
<td>40</td>
<td>83</td>
</tr>
<tr>
<td>$T_s$</td>
<td>CLR</td>
<td>5</td>
<td>39</td>
<td>83</td>
</tr>
<tr>
<td>$T_s$</td>
<td>$T_d$</td>
<td>0</td>
<td>2</td>
<td>99</td>
</tr>
</tbody>
</table>