Combination of Agglomerative and Sequential Clustering for Speaker Diarization

Deepu Vijayasenan, Fabio Valente and Hervé Bourlard
Idiap Research Institute CH-1920 Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Saketh Nath, Chiranjib Bhattacharyya
Indian Institute of Science Bangalore, India

Abstract
This paper aims at investigating the use of sequential clustering for speaker diarization. We had proposed a non-parametric method based on the agglomerative Information Bottleneck for fast diarization. Here we consider the combination of sequential and agglomerative clustering for avoiding local maxima of the objective function.

Introduction
Speaker diarization (who spoke when) involves unsupervised learning and model selection
Conventional diarization systems [1] are based on an ergodic HMM model
Agglomerative Clustering with a modified Bayesian Information Criterion for model selection
In our previous work, an alternative approach based on agglomerative Information Bottleneck had been proposed
Clusters segments based on their distance in a space of relevant variables
However, agglomerative clustering may converge to local optima of the objective function
In this work, we explore sequential clustering and combination of sequential clustering with agglomerative clustering to avoid any local optima.

Information Bottleneck (IB) Principle
Let X be set of elements to cluster
Y be the set of relevant variables
Relevance variables carry meaningful information about the problem
Choice of relevance variables depends on the domain. eg: Vocabulary of words can be considered as the relevance variables for document clustering
IB principle states that any compressed representation C of X should
minimize number of bits from X
keep meaningful information about Y
Optimal assignment is obtained by minimizing the following mutual information criterion w.r.t p(x|c),

\[ I(X, C) - \beta I(C, Y) \]  

where

\[ I(X, C) = \sum_{x \in X, c \in C} p(x) p(c|x) \log \frac{p(x|c)}{p(c)} \]  

\[ I(Y, C) = \sum_{y \in Y, c \in C} p(c) p(y|c) \log \frac{p(y|c)}{p(y)} \]  

IB Clustering
Agglomerative Information Bottleneck (aIB)
Greedy approach to minimize objective function of (1)
Algorithm is initialized with the trivial clustering of |X| clusters
Clusters are merged iteratively such that after each step the mutual information loss w.r.t the relevant variables is minimum
The mutual information loss \( \delta I_k \) in merging \( x_i \) and \( x_j \) is given by Jensen-Shannon divergence between \( p(y|x_i) \) and \( p(y|x_j) \):

\[ \delta I_k = (p(x_i) + p(x_j)) \cdot JS[p(y|x_i), p(y|x_j)] \]  

where \( JS \) denotes the Jensen-Shannon divergence.
Could converge to a local minimum

Sequential Information Bottleneck (sIB)
Finding the global minimum for given number of clusters (W)
 Initialized with a (possibly random) partition of W clusters
aIB draws a sample x at random, treats it as a singleton cluster and merges it to a new cluster \( c_{new} \) such that,

\[ c_{new} = \arg\min_{c \in C} JS(x, c) \]  

where \( JS(x, \cdot) \) is the Jensen-Shannon distance
At each step the objective function 1 improves or stays unchanged

Model Selection
Information theory based model selection:
Thresholding the normalized mutual information \( I(X, C)/I(C, Y) \)

Unsupervised Clustering
sIB operates with a fixed number of clusters. This number is not known a priori. Three methods of combination of aIB, sIB and model selection are investigated
aIB: Agglomerative clustering is performed and the best partition is chosen based on the model selection criterion
sIB: Sequential clustering is performed separately for each possible number of clusters. Model selection criterion then selects the best partition
aIB+sIB: aIB with model selection criterion is used to obtain an initial clustering which is then used as initialization to perform sIB

Speaker Diarization
Acoustic feature extraction from the audio file
Speech/non-speech segmentation and rejection of non-speech frames.
Uniform segmentation of speech in chunks of fixed size D
Estimation of GMM with shared diagonal covariance matrix i.e., definition of set Y
Estimation of conditional probability \( p(Y|X) \)
Clustering and model selection using one of the three methods
Viterbi realignment using conventional GMM system estimated from previous segmentation.

Experiments and Results
Performed all experiments on NIST RT06 eval data for “Meeting Recognition Diarization”
Speech/non-speech segmentation using a forced alignment of the reference transcripts on close talking microphone data
Pre-processing of the data consists of Wiener filtering for individual channels followed by a beam-forming algorithm (delay and sum). This was performed using the Beamformlt toolkit [2].

Following table presents the results of the three proposed methods, with oracle and Normalized Mutual Information model selection. The speaker error rates are reported before and after performing viterbi re-alignment.

<table>
<thead>
<tr>
<th>Model</th>
<th>aIB+</th>
<th>sIB</th>
<th>aIB+sIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>21.6</td>
<td>17.1</td>
<td>16.6</td>
</tr>
<tr>
<td>NMI</td>
<td>22.7</td>
<td>17.1</td>
<td>19.2</td>
</tr>
</tbody>
</table>

Speaker Error Rates for RT06 evaluation data. Same speech/non-speech segmentation as the baseline is used for all methods.

References