An overview of automatic speaker diarization systems

Wang Jun
CSLT, RIIT, THU

2012-10-27
Outline

1. Introduction to Speaker Diarization
2. General architecture of Speaker Diarization
3. Main approaches for speaker diarization
4. Brief Introduction of Algorithm
5. Comparison and Combination
6. Traditional Distance Metrics
7. Evaluation approach
8. Current Research Directions
9. outlook
Introduction to Speaker Diarization

- Speaker diarization is the task of determining "who spoke when?"

- Involve determining **the number of speakers** and identifying **the speech segments corresponding to each speaker**.

- A preprocessing for other downstream application. Such as speech retrieval, speech to text transcription and speaker recognition.
General architecture of Speaker Diarization

Figure 1 An overview of a typical diarization system
Main approaches for speaker diarization

Bottom-up approach:
- Training a number of clustering, merging and reducing the number of clusters until get the optimum number of clusters.

Top-down approach:
- Start with a single speaker model trained on all speech segment. Then add new speaker until the stop criterion.

Figure 2 Alternative clustering schemas
Brief Introduction of Algorithm

- Initialize clusters with the speech segments.
- Merge/split closest clusters.
- Update distances of remaining cluster to new cluster.
- Iterate until stopping criterion is met.
- Re-segmentation with GMM viterbi decoding.
Comparison and Combination

<table>
<thead>
<tr>
<th>Bottom-up approach</th>
<th>Top-down approach</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agglomerative hierarchical clustering.</td>
<td>Divisive hierarchical clustering.</td>
<td>Treat top-down output as a base segmentation and apply bottom-up output to purify it.</td>
</tr>
<tr>
<td>Use segment to train model is likely to capture more purer models. Bur it may corresponding to a single speaker or a phone class (short-term feature)</td>
<td>Use larger data to train small number of models Normalize both phone class and speaker. Can be purified.</td>
<td></td>
</tr>
</tbody>
</table>
Traditional Distance Metrics

0 The null hypothesis is that there is no speaker change at time $t$.

1 A speaker change point is hypothesized at time $t$. 

$$L_0 = \sum_{i=1}^{N_x} \log p(x_i | \theta_z) + \sum_{i=1}^{N_y} \log p(y_i | \theta_z)$$

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i | \theta_x) + \sum_{i=1}^{N_y} \log p(y_i | \theta_y).$$

LLR criterion: 

$$d_{\text{llr}} = L_1 - L_0.$$ 

BIC criterion: 

$$d_{\text{bic}} = L_1 - L_0 - \frac{\lambda}{2} \cdot \Delta K \cdot \log N.$$
Evaluation approach

- Dataset: NIST has organized a series of benchmark evaluations.
- Ground truth: manual labeling of acoustic data.
- DER is used as a results. It is composed as following figure.

DER = Speaker Error + False Alarm/Missed speech error + overlapped error

Large variations  Not robust  Stability SAD  Unsolved problem
Current Research Directions

- From features
  - time-delay features. Combine acoustic features and inter-channel delay feature.
  - Prosodic features in diarization.
  - Fusing short term and long term.

- From models
  - Use eigenvoice model to represent speaker.

- From metrics
  - Reference Speaker Model proposed by Wang Gang.
Current Research Directions

● New approaches
  ◆ the agglomerative information bottleneck (aIB)
  ◆ the sequential information bottleneck

To finding the most compact representation C of data X that minimizes the mutual information I(X,C) and preserves as much information as possible about Y (maximizing I(C, Y)). It can significant saving in computation.
Current Research Directions

◆ Bayesian machine learning
  not aim at estimating the parameters of a system (i.e. to perform point estimates), but rather the parameters of their related distribution (hyperparameters).

Bset model

\[ m = \arg\max_m p(m|Y) = \arg\max_m \frac{p(m)p(Y|m)}{p(Y)} \]

Marginal likelihood

\[ p(Y|m) = \int d\theta p(Y|\theta, m)p(\theta|m) \]

Traditional often use MAP to estimate parameter

\[ \theta_{MAP} = \arg\max_{\theta} p(\theta)p(Y|\theta) \]

BIC

\[ \log p(Y|m) = \log p(Y|m, \hat{\theta}) - \frac{\nu}{2} \log N \]

◆ Monte Carlo Markov Chains (MCMC) sampling method
Current Research Directions

- New approaches
  - Variational Bayes

\[
\log p(Y|m) = \log \int d\theta dX p(Y, X, \theta|m)
\]

Introduce a variational distribution and apply Jensen inequality to define the upper bound on the marginal log likelihood.

\[
\log p(Y|m) \geq \int d\theta dX \log q(X)q(\theta) \frac{p(Y, X, \theta|m)}{q(X)q(\theta)} = \\
= \int d\theta q(\theta) \int dX q(X) \log \frac{p(Y, X|\theta, m)}{q(X)} + \log \frac{p(\theta|m)}{q(\theta)} = \\
\int d\theta q(\theta) \int dX q(X) \log p(Y, X|\theta, m) - \int dX q(X) \log q(X) + \\
- \log \frac{q(\theta)}{p(\theta|m)} = F_m(q(X), q(\theta))
\]
outlook

- Overlapped speech.
- Robust to unseen variations.
- More efficient in order to process increasing dataset sizes.
- Aim at stream audio indexing.
References

Thanks