E-HMM approach for learning and adapting sound models for speaker indexing

Sylvain Meignier
Jean-François Bonastre
Stéphane Igounet
Outline

- Indexing and segmentation task
- Indexing and segmentation system
  - Overview
  - Indexing process
  - Example
- System parameters
- Database
- Results
  - NIST: n-segmentation task
  - Influences of background model and of learning parameters
Indexing and segmentation task

Indexing system must find the number of speakers, it must find who speaks and when

Hypotheses: the system do not know
- the number of speakers
- the speaker models

LIA Odyssey 2001
Indexing and segmentation system: Overview

- 2 processes
  - front-end
  - indexing

- Indexing process
  - iterative base on HMM
  - each state corresponds to a speaker model
  - adds a new speaker model
  - speaker models are adapted according to the segmentation at each stage

Front-end processing

Indexing process

- Adding a new speaker model
- Adapting speaker models
- Assessing the stop criterion

LIA Odyssey 2001
Indexing and segmentation system: example, stage 1

Stage 1: adding speaker L1

Process initialization

L1

L1
Indexing and segmentation system: example, stage 2

Stage 2: adding speaker L2

Process: steps 1 & 2

- The best subset is used to learn the L2 model, a new HMM is built.

Process: step 3, Models Adaptation

- According to the subset selected, this indexing is obtained.
  - No gain is observed, the adaptation of the L2 model is stopped.

Process: step 4, Stop criterion

- A gain is observed, a new speaker will be added.
  - Best 2 speakers indexing
  - Best one speaker indexing
Indexing and segmentation system: example, stage 3

Stage 3: adding speaker L3

Process: steps 1 & 2

The best subset is used to learn the L3 model, a new HMM is built.

Process: step 3, Models Adaptation

According to the subset selected, this indexing is obtained.

Process: step 4, Stop criterion

A gain is not observed, we return the best 2 speakers indexing.

No gain is observed, the adaptation of the L3 model is stopped.

LIA Odyssey 2001
**System parameters:**

**Front-end processing**

- Front-end processing
  - 16 cepstral coefficients + 16 delta coefficients
  - without frame removal
  - without CMS
  - using ELISA front-end processing

System parameters: Indexing process

Indexing process

- Speaker model adaptation
  - diagonal GMM 128
  - one pass of MAP
  - initialization with background model
    - ELISA development data (NIST evaluation 99 data)
    - composed of ~300 males and females (elec and carb)
  - only the means are adapted

Choice of the initial data of a new speaker model

- first adaptation is learned on 3 seconds of signal
- among the data labelled as L1
- we chose the 3 seconds with the maximum probability
**System parameters: Indexing process**

- **HMM transition probabilities**
  - let $n$ be the state number
  - probabilities to remain in the same state: $a=0.6$
  - probabilities to switch to another state: $b=(1-a)/(n-1)$

- **HMM emission probabilities**
  - likelihood ratios (using background model)
  - computed each 0.3 second

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>L2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>L3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6</td>
</tr>
</tbody>
</table>

LIA Odyssey 2001
Database

- CallHome corpus: NIST n-segmentation corpus
  - 500 telephonic conversions
  - From 2 up to 7 speakers per file
  - 6 languages

File number by speakers & by languages

<table>
<thead>
<tr>
<th>Speakers</th>
<th>File Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-sp</td>
<td></td>
</tr>
<tr>
<td>3-sp</td>
<td></td>
</tr>
<tr>
<td>4-sp</td>
<td></td>
</tr>
<tr>
<td>5-sp</td>
<td></td>
</tr>
<tr>
<td>6-sp</td>
<td></td>
</tr>
<tr>
<td>7-sp</td>
<td></td>
</tr>
</tbody>
</table>

Languages:
- Spanish
- Mandarin
- Arabic
- Japanese
- German
- English
Results:
NIST: n-segmentation task (1)

- 2 tests:
  - Trivial: scoring with only one segment
  - LIA: Nist 2001 result

- Results:
  - Has "good" results for each number of speakers
  - Is not optimized for speakers number
  - The language has "no" influence on the error rate

Nist Eval 2001 = 24%
Results:
Influence of background model and of learning parameters

- Computed on a half of the database (250 files)
- Different background models
  - SB: learned on SwitchBoard (ELISA-Dev corpus)
  - CH: learned on CallHome (NIST Dev database)
  - CH/SB: CallHome data adapted from SB model
    - CH/SB 0.7: MAP with 0.7 for SB model
- Different weight for MAP adaptation
  - Speaker model weight ∈ {0, 0.1, 0.2, 0.3, 0.4, 0.5}
Conclusion

- The result of n-segmentation is stable:
  - whatever the number of speakers is
  - whatever the speakers language is
- Our system is close to the best NIST 2001 system
  - LIA = 24%, Best = 20%
- Future development
  - for 2-segmentation, 2 speaker and tracking task
  - adding a duration model to the HMM