Blind Segmentation of a Multi-Speaker Conversation Using Two Different Sets of Features

Coby Metzger
A Monologue:
One person is talking to himself.

A Dialogue:
Two persons are talking to themselves.
The segmentation task

Phone → Segmentation → Training → Identification
Features Extraction

- FFT: 210-1510Hz  110Hz×12 +Δ×12
- Cepstra: LPC-C ×12 +Δ×12
Channel Compensation
Vox

FFT

CEPSTRA
Cluster Frames

2X3 SOM (Kohonen)

FFT
Model Each Cluster

CEPSTRA

3x10 SOM
Re-Label

\[ \log(P(f|0)) \quad \log(P(f|1)) \]

Smooth

Compare

Speaker 1
Speaker 0
Iterate

Re-Label

Log(P(f|0))  Log(P(f|1))

Smooth

Compare

Compute Likelihood

P(Component|Label)
Measuring Results - NIST
Measuring Results - NIST

True

Measured
The segmentation task

Phone

Segmentation

Training

Identification
Sensitivity

True

Measured
Specificity

True

Measured
Measuring Results - Revised

True

Measured

Sensitivity=0
Specificity=0.9
## Experimental Results

<table>
<thead>
<tr>
<th>Segment Length (sec)</th>
<th>Fraction of dominant speaker</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.74</td>
<td>0.49</td>
<td>0.66</td>
<td>0.76</td>
</tr>
<tr>
<td>40</td>
<td>0.72</td>
<td>0.54</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>60</td>
<td>0.69</td>
<td>0.68</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td>120</td>
<td>0.64</td>
<td>0.79</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td>300</td>
<td>0.61</td>
<td>0.83</td>
<td>0.86</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Summary

• Use of two different feature sets.
• An “engineering” approach to performance measure.
• Algorithm limitations:
  – More than 2 speakers.
  – Very unbalanced conversations.
  – Short conversations.
Demo