Audio Information Retrieval (AIR): Speaker ID (SID)

By Jacob Rettig
Mel-frequency cepstral coefficients

- MFCCs are commonly derived as follows:
- Take the **Fourier transform** of (a windowed excerpt of) a signal.
- Map the powers of the spectrum obtained above onto the **mel scale**, using **triangular overlapping windows**.
- Take the **logs** of the powers at each of the mel frequencies.
- Take the **discrete cosine transform** of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.
Segmenting Multimodal Streams

- Xerox PARC - Segmenting recorded meetings into categories.
- Used multi state hidden markov model (HMM), with each speaker having a subnetwork, the interconnections, and with a Gaussian output.
- Speech vectors from MFCCs (12)
- Speaker segmentation from a Viterbi decoder, noting the times when the optimal state sequence changes between sub networks
- HMM networks must be initialized. Done with agglomerative clustering and Baum Welch training algorithm. Iterative Viterbi decoding and training greatly improves results.

Table 3: Speaker Segmentation Error for Tied Gaussian Mixture.

<table>
<thead>
<tr>
<th></th>
<th>maximum</th>
<th>recomputed</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial</td>
<td>10.7%</td>
<td>4.2%</td>
</tr>
<tr>
<td>converged</td>
<td>.8%</td>
<td>.5%</td>
</tr>
</tbody>
</table>
Segmenting Multimedia Streams

- Foote - rapid speaker id using discrete MMI feature quantisation
- Uses 12 MFCC coefficients plus energy
- Uses supervised training and maximized mutual information (MMI) tree for quantisation rather than k-means. Better for high dimensional space.
- Tree created by splitting one coefficient dimension at a time
- Each threshold chosen to maximize the mutual information $I(X:C)$ between the data $X$ and the class labels $C$ that indicate the speaker generated
- pdf vectors created by following tree and counting elements in each cell divided by elements. Only top 3 to 303 due to silence
- Note all 13 coefficients useful rather than 2 for text. Due to nuances of vocal tract and pitch.