PSTL’s Speaker Diarization

Patrick Nguyen
Jean-Claude Junqua

Panasonic Speech Technology Laboratory (PSTL)
Plan

- System description
- SWAMP
System description

• Same as last year
• Segmentation: GMM decoding
• Five classes (narrow/wide x gender + sil)
• Heuristically smoothed

• + morph smoothing for zero-collar
System description

- Same as last year
- BIC with full covariance Gaussians

\[ \mathcal{N}(\mu_1, C_1) \quad \mathcal{N}(\mu_2, C_2) \quad \mathcal{N}(\mu_3, C_3) \quad \mathcal{N}(\mu_4, C_4) \]
High risk novel approaches

• Speaker diarization:
  – Closed form solutions
  – Quick test/train cycle
  – Few components: no interaction
  – Relatively new task
  – Self-contained
Frontend and Speaker recognition

- Standard question: why use STT features?
- No convincing alternative: why?
- Unlike linguistic content, speaker characteristics do not show on spectrogram
- Unknown distortions in the spectrogram contain the information

=> Purely mathematical signal processing
SWAMP

- SWAMP = Sweeping metric parameterization
- Input: MFCC
- Output: Ranked SWAMP features

- Basic idea: speech is a time curve on a manifold
Metaphors: surface and grid

$C^\infty$ manifold: compact $E$-object.

$(E + 1)$ embedding

projected
Why Isometric Features?

• Simple algebra and Gaussian OK

\[ \mu = \frac{1}{2} [x_1 + x_2] \]

SWAMP: isometric

MFCC
Sweeping metric

- Defines the shape of the squares [Scroedinger, Levin]

\[ g^{kj} = \left\langle \frac{\partial x^k}{\partial t} \frac{\partial x^j}{\partial t} \right\rangle_{x \approx y} = g^{jk} \]

- Intuition: fast is imprecise

Defined locally around y
Sweeping metric(2)

- Each square has its metric $g$
Idea: triangulation

- Characterize all points by relative distance
- Flatten out surface

\[ \Delta(a, b) = \int_a^b ds^2 \]

\[ ds^2 = g_{kj} dx^k dx^j \]

“Geodesic lengths”
Classical Metric Multidimensional Scaling (CMDS)

- Visualization technique used in psychology

<table>
<thead>
<tr>
<th></th>
<th>mi</th>
<th>SB</th>
<th>Boston</th>
<th>NYC</th>
<th>Dallas</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB</td>
<td>0</td>
<td>2663</td>
<td>2521</td>
<td>1326</td>
<td></td>
</tr>
<tr>
<td>Boston</td>
<td>0</td>
<td>190</td>
<td></td>
<td>1551</td>
<td></td>
</tr>
<tr>
<td>NYC</td>
<td>0</td>
<td></td>
<td>1373</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dallas</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Isometric coordinates
Distances and MDS

• Summary:
  – Find local metrics $g^{kj}$
  – Find geolen distances by integration
    \[ \Delta = \int ds^2 \]
  – SVD the distance matrix (MDS)

<table>
<thead>
<tr>
<th></th>
<th>SB</th>
<th>Boston</th>
<th>NYC</th>
<th>Dallas</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB</td>
<td>0</td>
<td>2663</td>
<td>2521</td>
<td>1326</td>
</tr>
<tr>
<td>Boston</td>
<td>0</td>
<td>190</td>
<td>1551</td>
<td></td>
</tr>
<tr>
<td>NYC</td>
<td></td>
<td></td>
<td>0</td>
<td>1373</td>
</tr>
<tr>
<td>Dallas</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Implementation details

- Vector quantization: curse of dimensionality
- Voronoi regions: locally linear regions
- Adjacency: join squares of the surface
- Tunneling: accidentally go “under” surface
- Floyd-Warshall: complete paths
Vector Quantization

• Curse of dimensionality prevents from computing local metrics (space not compactly populated)

• Alternative (e.g. Levin): Riemann-Christoffel symbol of the second kind

\[ \Gamma^{kj}_{\ i} \propto \left( \frac{\partial g^{kj}}{\partial x^i} \right) \]

• Large ODE
Vector Quantization

• Our solution: locally linear regions
• Simplifies the geolen computation as well
• Problem: VQ distortion is initialized with the contravariant coordinates

\[ g^{kj}(1) \]
\[ g^{kj}(2) \]
\[ g^{kj}(3) \]
Voronoi regions

- Polyhedral geodesics are straight lines if local embeddings are linear
- Problem: Dirichlet/Voronoi interface is quadratic
- Newton/Raphson solution of the Lagrangian

Voronoi interface
Floyd-Warshall (or Dijkstra)

- From local geolens to complete graph
- Local geolens: Completed geolens
Tunneling and Adjacency

• Use time structure to avoid going under the manifold
Properties

• Unique under any invertible constant distortion
• Robust to quasi-stationary noise
• Isometric
• Time-distortion can be recovered by extending the vector (WRONG?)
• Expensive
Future work?

• Probabilistic measure? Fisher?
• Non-Riemannian topology?
  – Information divergence [Amari]
  – Renyi cross entropy?
• Spectral energy?
• ASR?
## Are segmentations created equal?

<table>
<thead>
<tr>
<th>Seg/Cluster</th>
<th>Spk Error</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU/CU</td>
<td>63.04%</td>
<td></td>
</tr>
<tr>
<td>CU/PSTL</td>
<td>18.51%</td>
<td>17.85%</td>
</tr>
<tr>
<td>BBN/BBN</td>
<td>~27%</td>
<td></td>
</tr>
<tr>
<td>BBN/PSTL</td>
<td>19.33%</td>
<td>17.91%</td>
</tr>
<tr>
<td>CLIPS/CLIPS</td>
<td>57.91%</td>
<td></td>
</tr>
<tr>
<td>CLIPS/PSTL</td>
<td>30.24%</td>
<td></td>
</tr>
<tr>
<td>LIA/LIA</td>
<td>19.21%</td>
<td>18.46%(ad)</td>
</tr>
<tr>
<td>LIA/PSTL</td>
<td>16.14%</td>
<td></td>
</tr>
<tr>
<td>LIMSI/LIMSI</td>
<td>35.10%</td>
<td></td>
</tr>
<tr>
<td>LIMSI/PSTL</td>
<td>19.03%</td>
<td>17.56%</td>
</tr>
<tr>
<td>MITLL2/MIT</td>
<td>20.04%</td>
<td></td>
</tr>
<tr>
<td>MITLL2/PSTL</td>
<td>18.59%</td>
<td></td>
</tr>
<tr>
<td>PSTL/PSTL</td>
<td>18.58%</td>
<td>17.94%</td>
</tr>
</tbody>
</table>
Zero Collar

- Zero collar requires precise speech/non-speech
- Morph filter minimizes the risk (trim a little from both ends)
- Apply 0.3s bridging rule

- (old system): 19.33% spkr error => 18.71%
Results

• Zero collar is sharp
• Length of shows affects performance
• Last year: 18.8% spkr time err
• This year: 13.5% (18.6% relative)
• Only participant (PSTL) last year
• Lowest error
Conclusion

• Novel approach (SWAMP)
  – Sweeping metric
  – Geolen computation
  – Multidimensional scaling

• Simple full covariance BIC clusterer