Introduction

- Speaker Diarization: “Who Spoke When?”
- Unsupervised speaker diarization.
- Uses "Remember–Learn–Transfer" principle to transfer the learned information.
- Reduces the real time factor of two-pass system.

IB based Diarization

- A set of segments $\mathcal{X}$ in an audio is clustered into set of clusters $\mathcal{C}$ preserving the relevant information $\mathcal{Y}$. The objective function is given by $F = I(\mathcal{Y}, \mathcal{C}) - \frac{1}{\beta} I(\mathcal{C}, \mathcal{X})$

Two-pass IB based diarization

- First pass: IB based diarization is performed to obtain relative speaker labels.
- ANN Training & LSF Extraction: ANN initialized with random weights is trained from scratch on the output boundary labels and the spectral features to obtain latent features (LSF).
- Second pass: The LSFs are used along with the spectral features in the second pass of IB system.

TPIB-ITL

- Training seedANN: A seedANN is trained from the first audio to be diarized by the system.
- First pass: IB based diarization is performed to obtain relative speaker labels.
- Remember–Learn–Transfer: ANN initialized with weights from seedANN is fine-tuned on the output boundary labels and the spectral features of the current recording to obtain LSF. Store the fine-tuned ANN for next recording.
- Second pass: A second pass of IB based clustering is performed.

Incremental Transfer Learning in TPIB

- First Pass: IB based diarization is performed to obtain relative speaker labels.
- Orthogonalization
- Posterior Estimation (PS)
- KL-HMM Realignment
- First Pass Output
- Second Pass: The LSFs are used along with the spectral features in the second pass of IB system.
- TPIB-ITL: Incremental Transfer Learning in TPIB

Important Result

- TPIB-ITL uses "Remember–Learn–Transfer" principle to diarize new recordings.
- Retaining previous knowledge helps to reduce real time factor compared to TPIB system.

Speaker Error Rates

<table>
<thead>
<tr>
<th>System</th>
<th>Feature(s)</th>
<th>Dev. Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RT-04Dev</td>
<td>RT-04Eval</td>
</tr>
<tr>
<td>IB</td>
<td>MFCC</td>
<td>15.1</td>
<td>13.5</td>
</tr>
<tr>
<td>TPIB</td>
<td>MFCC+LSF (0.8, 0.2)</td>
<td>13.1</td>
<td>12.5</td>
</tr>
<tr>
<td>TPIB-ITL</td>
<td>MFCC+LSF (Avg.)</td>
<td>14.9</td>
<td>12.6</td>
</tr>
<tr>
<td>Proposed System</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPIB-ITL</td>
<td>LSF</td>
<td>15.5</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>MFCC+LSF (0.1, 0.9)</td>
<td>15.2</td>
<td>12.2</td>
</tr>
<tr>
<td></td>
<td>MFCC+LSF (Avg.)</td>
<td>15.8</td>
<td>12.5</td>
</tr>
<tr>
<td>TPIB-ITL</td>
<td>LSF</td>
<td>15.5</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>MFCC+LSF (0.1, 0.9)</td>
<td>15.2</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>MFCC+LSF (Avg.)</td>
<td>15.8</td>
<td>13.3</td>
</tr>
</tbody>
</table>

Real Time Factors

Overall RTF

- TPIB-ITL also works when only development data in used in incremental learning phase.

Conclusion

- No separate training data is used.
- Recording-specific discrimination is achieved.
- Sequence of recordings does not affect the performance.
- TPIB-ITL also works when only development data in used in incremental learning phase.

References