1. **Speaker Change Detection**
   - We are interested in the time points at which the change happen.
   - Speaker identity is not important.
   - Existing methods are based on comparing the features from two consecutive segments.
   - For an online application, can we operate with segments of at most 2 seconds?

2. **Method**
   1. Gender classification
   2. Contrastive loss training
   3. Triplet loss training

   \[ L_e = \sum_{m=1}^{M} \delta(s_i^{(m)} = s(m))d(x_i^{(m)}, x_j^{(m)}) + \delta(s_i^{(m)} \neq s(m)) \max(0, \Delta_e - d(x_i^{(m)}, x_j^{(m)})) \]  
   \[ L_{tr} = \sum_{m=1}^{M} \max(0, \Delta_{tr} + d(x_i^{(m)}, x_j^{(m)}) - d(x_i^{(m)}, x_k^{(m)})) \]

3. **Results: Accuracy**
   - Dataset:
     - 144 hours of audio from LDC HUB4 Broadcast News.
     - Training segments have duration of 2s.
     - Sampled 500k pairs, 329 triplets for training.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>52.2</td>
</tr>
<tr>
<td>i-vector</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>86.6</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Gender</td>
<td>Yes</td>
<td>76.9</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Gender</td>
<td>No</td>
<td>78.1</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Contrast</td>
<td>Yes</td>
<td>77.4</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Contrast</td>
<td>No</td>
<td>87.5</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Triplet</td>
<td>Yes</td>
<td>82.7</td>
</tr>
<tr>
<td>PLP</td>
<td>S+C</td>
<td>Triplet</td>
<td>No</td>
<td>89.0</td>
</tr>
</tbody>
</table>

   Results:
   - Among three pre-training methods triplet loss is the best.
   - Using Euclidean distance is slightly better than the cosine distance.

4. **Test Setup**
   - 10 audio files are chosen for test.
   - Left-right comparisons are performed around segment boundaries rather than using sliding windows for low-latency.
   - Choice of segments:
     - Based on segment type
       1. ASR
       2. Ground truth
     - Based on segment duration
       1. Variable length (< 2s)
       2. Fixed length

5. **Results: Precision-Recall and F-Measure**
   - (a) Variable length-ASR
   - (b) 2s-ASR
   - (c) Variable length-Ground truth
   - (d) 2s-Ground truth

<table>
<thead>
<tr>
<th>ASR boundary</th>
<th>Ground truth boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>2-second</td>
</tr>
</tbody>
</table>
   | i-vector     | 0.3150                | 0.4902       | 0.5036                | 0.6109
   | Tri-Eucl-F   | 0.3332                | 0.4591       | 0.4722                | 0.5736
   | Tri-Eucl-T   | 0.4746                | 0.5323       | 0.6141                | 0.6511

   - Relative improvements in F-measures as compared to i-vectors are:
     - 50.7% in the highly mismatched condition (ASR-Variable length)
     - 6.6% in the matched condition (Ground truth-2s)

   - Score combination of i-vector and Triplet-T system performs 5% better on 2s segments.

6. **Conclusions**
   1. Jointly trained Siamese network and the classifier performs better than classifying i-vectors.
   2. Siamese embeddings are more robust to the duration mismatch between training and test segments.
   3. Siamese embeddings perform better than i-vectors for ≤2s segments which is important for achieving low-latency.