Sufficiency Quantiﬁcation for Seamless Text-Independent Speaker Enrollment

Gokcen Cilingir, Jonathan Huang, Mandar S Joshi, Narayan Biswal
Intel Corporation

**Motivation**
- **NOW**
  - Voice-enabled platforms are taking off
  - Text-dependent speaker recognition is a technology that is already in the market in smart speakers
- **FUTURE**
  - Speaker recognition (SR) on natural speech (Text-independent speaker recognition)
- **PROBLEM**
  - Enrolling for SR with natural speech is NOT user-friendly
- **SOLUTION**
  - Seamless enrollment/adaptation

**Phoneme-richness quantiﬁcation – oﬄine process**

- Data: corpus
- VAD
- UBM-modeling with GMM
- UBM
- Universal background model (UBM) models speech
- Phoneme-richness quantiﬁcation over enrollment utterance pool
- UBM
- Phoneme-richness score of the enrollment utterance pool
- Reference histogram represents the ground truth for an adequately phoneme-rich utterance pool
- Reference histogram
- Phoneme richness histogram
- Utterance richness histogram
- Phoneme richness score
- Score of the enrollment utterance pool
- Adapt-create speaker model
- Adapt-creating speaker model
- Utterance richness histogram
- Utterance pool
- Catalog utterance with Speaker ID
- Utterance

**Experiment Results**

**Enrollment scenarios**

<table>
<thead>
<tr>
<th>Enrollment scenario</th>
<th>Speech duration (seconds)</th>
<th>Phoneme-richness score (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 &quot;hello computer&quot; repeats</td>
<td>0.102</td>
<td>0.152</td>
<td>37.44</td>
</tr>
<tr>
<td>10 repeats of 4 trigger words</td>
<td>13.32</td>
<td>0.120</td>
<td>30.12</td>
</tr>
<tr>
<td>Short phoneme-rich passage</td>
<td>12.35</td>
<td>0.152</td>
<td>23.56</td>
</tr>
</tbody>
</table>

**Sufficiency metrics**

- Success criteria (Pearson correlation with EER)
- Phoneme-richness score
- Speech duration

**Phoneme-richness quantiﬁcation usages beyond seamless enrollment**

- Improve UX for traditional text-independent SR enrollment
- Conﬁdence modeling during detection/test.
- For multi-session enrollment and model fusion

**Future Direction**

- Proving metric utility over low quality data, spontaneous and noisy speech
- Proving metric utility over conﬁdence modeling task

**Data set**

- A proprietary dataset of 40 speakers.
- Each speaker uttered 180 short commands, 10 repeats of 20 trigger words and a short phoneme-rich passage.
- 180 short commands are split into 25 batches, where each batch contained ~3 secs of speech content.
- 10 batches left out for testing. Equal error rate (EER) is an accuracy metric.

**Sufficiency quantification**

- What criteria will you use to decide that you have enough data in the utterance pool to create a phrase-independent model of the speaker?
  - Current enrollment model: Ask user to read a phoneme-rich passage, or a couple of such sentences
  - Naive approach: Use speech duration
  - Our approach: Deﬁne a metric to quantify phoneme-richness

**Experiment Results**

<table>
<thead>
<tr>
<th>Enrollment scenarios</th>
<th>Speech duration (seconds)</th>
<th>Phoneme-richness score (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 &quot;hello computer&quot; repeats</td>
<td>0.102</td>
<td>0.152</td>
<td>37.44</td>
</tr>
<tr>
<td>10 repeats of 4 trigger words</td>
<td>13.32</td>
<td>0.120</td>
<td>30.12</td>
</tr>
<tr>
<td>Short phoneme-rich passage</td>
<td>12.35</td>
<td>0.152</td>
<td>23.56</td>
</tr>
</tbody>
</table>

**Sufficiency metrics**

- Success criteria (Pearson correlation with EER)
- Phoneme-richness score
- Speech duration

**Future Direction**

- Proving metric utility over low quality data, spontaneous and noisy speech
- Proving metric utility over conﬁdence modeling task