Personalized Acoustic Modeling by Weakly Supervised Multi-task Deep Learning Using Acoustic Tokens Discovered from Unlabeled Data
Cheng-Kuan Wei, Cheng-Tao Chung, Hung-Yi Lee, and Lin-Shan Lee
National Taiwan University

1. Introduction
- Motivation: with the popularity of smartphones,
  - Easier to collect huge personal audio data which can be used
  - Little transcribed data as references for learning
- Proposed: phoneme-token deep neural network (PTDNN)
  - Jointly trained by unsupervised acoustic tokens + transcribed phonemes
  - "Weakly supervised" scenario
  - Vast unlabeled data & limited transcribed data

2. Proposed Approach
- Acoustically similar signal patterns automatically discovered from a large unlabeled corpus [1]
- Specified by number of token HMM states \( m \) (token length or temporal granularity) and number of distinct tokens \( n \) (phonetic granularity)
- Offer high level information regarding how each signal segment sounds like
- Use token HMM states together with phoneme HMM states as learning targets:
  \[ \text{Token states} \quad \bullet \quad \bullet \quad \bullet \quad \text{Token } i \quad \text{Token } j \]

3. Experiments & Analysis
- SI training data: 31.8 hours Chinese + 29.7 hours English read speech
- Adaptation data: Facebook posts produced by 5 male, 5 female
- Chinese-English Bilingual (4.1%: English words)
- 1000 utterances for each of 10 speakers
- 500 Adapt. (M transcribed + (500 – M) unlabeled), 250 Valid., 250 Testing
- Personal acoustic token sets: train on adaptation set of each speaker

- 50 transcribed + 450 unlabeled utterances (from adaptation set)
- Robust to the choice of \( m \) and \( n \)

- More or Less Transcribed Data and More Token Sets

- Word accuracy

<table>
<thead>
<tr>
<th>Models</th>
<th>Number of transcribed utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI (DNN-HMM)</td>
<td>10</td>
</tr>
<tr>
<td>(A)</td>
<td>59.3%</td>
</tr>
<tr>
<td>(B-1)</td>
<td>60.3%</td>
</tr>
<tr>
<td>(B-2)</td>
<td>60.34%</td>
</tr>
<tr>
<td>(B-3)</td>
<td>61.45%</td>
</tr>
<tr>
<td>(C)</td>
<td>64.18%</td>
</tr>
</tbody>
</table>