1. Introduction

Developmental Speech Sound Disorders (SSDs) are a common communication impairment in childhood that have the potential to negatively affect the lives and the development of children. Clinical intervention is typically available for children with SSDs, but current clinical methods for speech therapy are subjective and time consuming.

In the Ultrax Speech Project, we explore objective methods that could alleviate manual processes undertaken by Speech and Language Therapists (SLTs) using audio and ultrasound.

2. The Ultrasurete Repository

UltraSuite is a repository of ultrasound and acoustic data from child speech therapy sessions [1]. This repository contains three separate datasets, one of typically developing (TD) children and two of children with speech sound disorders (SSD).

The two SSD datasets are divided into assessment and therapy sessions. Assessment sessions include:
- Baseline - BL
- Mid-Therapy - Mid
- Post-Therapy - Post
- Maintenance - Maint

The two SSD datasets are divided into assessment and therapy sessions. Assessment sessions include:
- Baseline - BL
- Mid-Therapy - Mid
- Post-Therapy - Post
- Maintenance - Maint

3. Main Challenges

- Interaction between therapist and child.
- Mispronunciations
- Insertions and deletions with respect to the given prompt.
- Child speech processing
- Disordered speech processing

4. Speaker Labelling

Transcriptions (available only for the UXTD dataset) were reduced to CHILD and SLT tokens. These were modelled with 5-state ergodic HMMs. Silences were modelled with 5 state left-to-right skip HMMs.

5. Word Alignment

Robust word alignment is of particular importance to alleviate the manual steps taken by SLTs. This involves time-aligning relevant keywords, suggested by the prompt, with the speech recording.

We begin by building various baselines to illustrate the main challenges. Results on Table 2 illustrate the impact of the speaker labelling model. Results on Table 3 investigate additional training data of child speech.

5. Word Alignment

<table>
<thead>
<tr>
<th>Speaker labels</th>
<th>Word scoring</th>
<th>Time scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td>no no</td>
<td>0.482</td>
<td>0.475</td>
</tr>
<tr>
<td>no yes</td>
<td>0.533</td>
<td>0.517</td>
</tr>
<tr>
<td>yes yes</td>
<td>0.467</td>
<td>0.460</td>
</tr>
<tr>
<td>yes yes</td>
<td>0.577</td>
<td>0.566</td>
</tr>
</tbody>
</table>

Table 2: Effect of removing SLT time segments from speaker labelling model. Averaged results (TD, SSD) from HMM-DNN trained on UXTD and UPX.

Precision and Recall are measured on retrieved word boundaries (allowing a 10ms collar) as well as retrieved time segments (in seconds).

6. Future Work

Baseline systems show that there is plenty of room for improvement, especially with SSD data (Table 4).

Going forward:
- Acoustic modelling: out-of-domain data, transfer learning
- Speaker-dependent pronunciation modelling
- Ultrasound data
- Insertions, deletions, and deviations from prompt.