Towards Robust Word Alignment of Child Speech Therapy Sessions

Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, Steve Renals
Centre for Speech Technology Research, University of Edinburgh, UK
sam.ribeiro@ed.ac.uk

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 UltraSuite 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 are:

- Baseline - BL
- Mid-Therapy - Mid
- Post-Therapy - Post
- Maintenance - Maint

	UXTD	UXSSD	UPX
Speakers	58	8	20
Gender (M/F)	27/31	6/2	16/4
Age	5-12	5-10	6-13
Total speech (hrs)	3.47	5.47	9.19
Child (hrs)	2.24	3.66	7.27
SLT (hrs)	1.24	1.81	1.92
Total silence (hrs)	4.40	5.16	9.59
Total audio (hrs)	7.87	10.63	18.78

Table 1: UltraSuite repository, with hours of speech and silence.

[1] Aciel Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James Scobbie, and Alan Wrench. Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *INTERSPEECH*, Hyderabad, India, 2018.

3. Main challenges

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

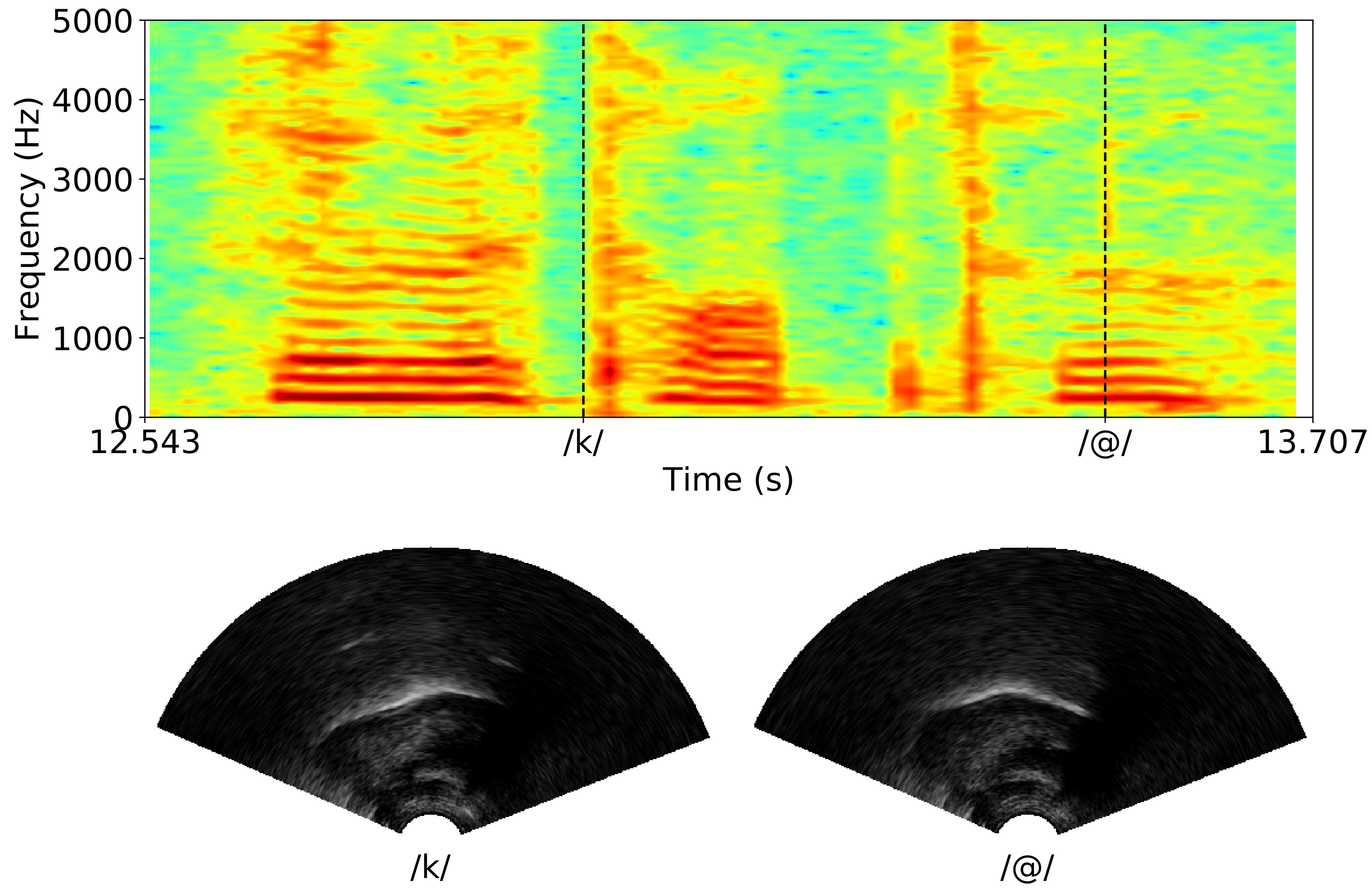

Figure 1: Spectrogram for the word *helicopter* with two corresponding ultrasound frames, ellicited during a session with a six-year-old child diagnosed with velar fronting. Ultrasound frames show a mid-saggital view of the oral cavity with the tip of the tongue facing right.

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.

Force-aligned transcriptions from held-out TD data were used as a ground truth. Identification Error Rate (IER), precision, and recall were measured in terms of seconds.

IER: **4.6%** Precision: **0.969** Recall: **0.979**

The three datasets were decoded using this method, which formed the basis for the estimates reported in Table 1.

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.

Training data	Model	Word scoring Prec	Word scoring Rec	Word scoring f1	Time scoring Prec	Time scoring Rec	Time scoring f1
UXTD	GMM	0.327	0.147	0.171	0.638	0.201	0.258
UXTD, UPX	GMM	0.604	0.585	0.594	0.758	0.692	0.722
	DNN	0.577	0.566	0.571	0.700	0.700	0.700
UXTD, UPX, PFSTAR	GMM	0.646	0.632	0.639	0.786	0.738	0.760
	DNN	0.654	0.642	0.648	0.774	0.760	0.765
UXTD, UPX, OGI	GMM	0.564	0.552	0.558	0.718	0.667	0.691
	DNN	0.602	0.590	0.596	0.737	0.731	0.733
UXTD, UPX, PFSTAR, OGI	GMM	0.566	0.554	0.560	0.713	0.655	0.681
	DNN	0.610	0.598	0.604	0.726	0.713	0.718

Table 3: Averaged results (TD, SSD) on additional training data.

Speaker labels Train	Speaker labels Test	Word scoring Prec	Word scoring Rec	Word scoring f1	Time scoring Prec	Time scoring Rec	Time scoring f1
no	no	0.482	0.475	0.478	0.614	0.606	0.608
no	yes	0.533	0.517	0.525	0.622	0.631	0.625
yes	no	0.467	0.460	0.463	0.567	0.577	0.571
yes	yes	0.577	0.566	0.571	0.700	0.700	0.700

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 100ms collar) as well as retrieved time segments (in seconds).

Additional child speech data:

- PF-STAR corpus: 7.5hrs, 86 children, BrE
- OGI corpus: 22.5 hrs, 500 children, AmE

Systems:

- GMM: Triphone model with LDA, MLLT, SAT.
- DNN: Feedforward network with 6 layers and RBM pre-training *(nnet1)*.

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.

Dataset	Subset	Word scoring Prec	Word scoring Rec	Word scoring f1	Time scoring Prec	Time scoring Rec	Time scoring f1
UXSSD	BL	0.524	0.504	0.513	0.766	0.673	0.716
	Mid	0.713	0.687	0.700	0.788	0.746	0.766
	Post	0.625	0.605	0.615	0.759	0.711	0.735
	Maint	0.572	0.572	0.572	0.679	0.647	0.662
	mean	*0.609*	*0.592*	*0.600*	*0.748*	*0.694*	*0.720*
UXTD	dev	0.737	0.737	0.737	0.802	0.892	0.845
	test	0.754	0.745	0.749	0.848	0.890	0.868
	mean	*0.746*	*0.741*	*0.743*	*0.825*	*0.891*	*0.857*

Table 4: Results per evaluation set for best baseline system (DNN trained on UXTD, UPX, PFSTAR).