UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions

Citation for published version:

Digital Object Identifier (DOI):
10.21437/Interspeech.2018-1736

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

We introduce UltraSuite, a curated repository of ultrasound and acoustic data, collected from recordings of child speech therapy sessions. This release includes three data collections, one from typically developing children and two from children with speech sound disorders. In addition, it includes a set of annotations, some manual and some automatically produced, and software tools to process, transform and visualise the data. The repository will continue to grow and become larger and more comprehensive as we add new studies, ensuring that all new data is available in the same standardised format. We invite other researchers to contribute their data to this repository.

1. Introduction

Speech sound disorders (SSDs) affect quality of life for a large number of children. In the UK, 11.4% of eight year olds have persistent SSDs, ranging from common clinical distortions to speech that is unintelligible even to close family members [1]. SSDs are similarly prevalent in other countries [1]. Children with disordered speech experience adverse outcomes of many kinds: social and psychological outcomes, difficulty with literacy and educational attainment, and long-term employment prospects [2, 3, 4]. However, current clinical practice for assessing these disorders is subjective and inaccurate [6]. Instrumental methods that use articulatory imaging, such as ultrasound, provide a more accurate diagnosis, at the expense of large amounts of manual effort from a highly trained pathologist. Machine learning has the potential to automate much of this work, leading to better outcomes for patients without increasing workload for pathologists, but publicly available data that could facilitate this work is scarce. Existing work reports results on adult data [7, 8, 9], data that is from the average tongue shape in correct articulation. While ultrasound imaging of the tongue provides additional information not available in the acoustic signal (e.g., the presence of double articulations or undifferentiated lingual gestures [10]), this additional information reduces subjectivity and in some cases changes the diagnosis. However, working with speech recordings and ultrasound videos is time consuming and difficult.

In order to provide a diagnosis or measure therapy progress, the speech pathologist goes through the following process: searching therapy recordings for occurrences of words of interest (the recordings often contain background noise, the voice of the therapist encouraging the child to speak, and the child uttering multiple words and potentially making mistakes); identifying boundaries of a phone of interest both in the audio and ultrasound video; locating the mid-frame of the phone in the ultrasound video; fitting a contour to the tongue in the image; and finally, measuring how similar/dissimilar the tongue shape is from the average tongue shape in correct articulation. Working with speech recordings and ultrasound videos is time consuming and difficult. Automating it would allow it to be offered to children as standard practice.

1.2. Broader Applicability of the Data

Access to the articulatory domain through imaging techniques such as ultrasound gives additional information over the acoustic domain. Indeed, in addition to the area of speech and language pathology, prior work has shown that articulatory information has the potential to improve performance in multiple aspects of speech technology, for instance: speech recognition [19], speech synthesis [20], and silent speech interfaces [21].
Table 1: The number of participants, their gender and ages. We report ages in years (y) and months (m). We recorded the ages of participants on their single visit in UXTD, and on their first baseline in UXSSD and UPX.

<table>
<thead>
<tr>
<th>UXTD</th>
<th>UXSSD</th>
<th>UPX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>58</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>Male</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>Mean age</td>
<td>9y 3m</td>
<td>7y 7m</td>
</tr>
<tr>
<td>SD age</td>
<td>1y 10m</td>
<td>1y 6m</td>
</tr>
<tr>
<td>Min age</td>
<td>5y 8m</td>
<td>5y 11m</td>
</tr>
<tr>
<td>Max age</td>
<td>12y 10m</td>
<td>10y 1m</td>
</tr>
</tbody>
</table>

2.1. Typically Developing Children

We recorded the typically developing subset of the Ultrax dataset (UXTD) between 11/2011–10/2012. The purpose of the experiment was to evaluate the effectiveness of ultrasound as a visual biofeedback tool for learning non-English articulations. Each child attended once and recorded a single session.

2.2. Children with Speech Sound Disorders

The repository at this stage contains two datasets recorded with children with speech sound disorders. The first is the Ultrax speech sound disorders subset (UXSSD) which we recorded between 12/2011–07/2014, and the second is the UltraPhonix dataset (UPX), recorded between 06/2015–03/2017. The children exhibited a range of SSDs including phonological delay, phonological disorder, inconsistent phonological disorder, vowel disorder, articulation disorder, and childhood apraxia of speech. The data was recorded specifically for the purpose of evaluating the effectiveness of ultrasound as a visual biofeedback tool for therapy. Each child attended several sessions: suitability (before baseline), baseline (1–5 sessions), therapy (1–12 sessions), mid-therapy, post-therapy (immediately after therapy), and maintenance (several months after therapy). Table 1 shows the number of participants in each of the three datasets, their gender and ages. Persistent SSDs are more commonly associated with boys than girls, which explains the gender imbalance in the data.

3. Data Preparation

We exported the raw data from the proprietary AAA format to obtain a tuple of four files per utterance:

1. **Prompt file**: contains text describing the task the child was given and the date-time of recording.
2. **Audio file**: RIFF wave file, sampled at 22.05 KHz, containing the speech of the child and the SLT.
3. **Ultrasound file**: a sequence of ultrasound frames capturing the midsagittal view of the child’s tongue. A single ultrasound frame is recorded as a 2D matrix where each column represents the ultrasound reflection intensities along a single scanline. The surface of the probe is convex and the scanlines are directed in an equal-angled fan in the scanning plane. In order to correctly interpret the ultrasound data, a set of parameters are recorded in the parameter file described below.
4. **Parameter file**: contains a set of parameters to interpret the ultrasound data and synchronise it with the audio. It gives the number of scanlines in each frame (63), the number of data points per scanline (412), number of bits used to represent each reflection intensity data point (8), the angle between each scanline (0.038°), and the synchronization offset relative to the audio in seconds.

We discarded utterance tuples where the audio was too short and was unrelated to the prompt.

3.1. Prompts

We standardised the formatting of the prompt text by removing inconsistencies, such as replacing tabs with white spaces, removing duplicate or trailing white spaces, correcting the capitalisation of proper nouns, and correcting misspellings. We identified six distinct types of prompts:

A. **Words**: a group of semantically unrelated English words (e.g., “down link pat get”) which were either identified by the SLT as being diagnostically useful, or were based on a protocol from the Diagnostic Evaluation of Articulation and Phonology (DEAP).

B. **Non-words**: designed to elicit certain phones from the child but which are not real words (e.g., “\d apa epe opo\d”).

2We excluded from the repository participants whose guardians did not provide consent.
Table 2: The number of utterances per prompt type, with the number of unique prompts in parentheses for each of the three datasets. We encode the type identifier in the file names.

<table>
<thead>
<tr>
<th>Type</th>
<th>ID</th>
<th>UXTD</th>
<th>UX SSD</th>
<th>UPX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>A</td>
<td>962 (26)</td>
<td>2708 (291)</td>
<td>3838 (455)</td>
</tr>
<tr>
<td>Non-words</td>
<td>B</td>
<td>607 (27)</td>
<td>495 (59 )</td>
<td>560 (60)</td>
</tr>
<tr>
<td>Sentence</td>
<td>C</td>
<td>0 (0)</td>
<td>445 (35)</td>
<td>1020 (128)</td>
</tr>
<tr>
<td>Articulatory</td>
<td>D</td>
<td>2934 (45)</td>
<td>132 (17)</td>
<td>211 (31)</td>
</tr>
<tr>
<td>Non-speech</td>
<td>E</td>
<td>116 (2 )</td>
<td>9 (1)</td>
<td>302 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>F</td>
<td>0 (0)</td>
<td>56 (12)</td>
<td>61 (7)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>4619 (100)</td>
<td>3845 (415)</td>
<td>5992 (682)</td>
</tr>
</tbody>
</table>

(C) Sentence: designed to elicit co-articulation (e.g., “It’s a toe Pam”) or designed to examine phones of interest in different contexts at the sentence level (e.g., “My Granny Maggie got a golden gown” where /g/ occurs in different word positions and different vowel environments).

(D) Articulatory: single or multiple phones occurring once or repeated. The SLT pronounces a phone, or plays a recording of a phone at different speeds, and the child is expected to imitate what they hear. The latter is known as a Diadochokinesis imitation task [23].

(E) Non-speech: includes swallowing motions recorded to obtain a trace of the hard palate, and coughs recorded to obtain additional tongue shapes.

(F) Other: conversational speech, such as describing a picture or telling a story about it (e.g., “Connected speech picture 1”).

The number of utterances per prompt type in each dataset and the number of unique prompts are shown in Table 2.

3.2. File Naming Convention

We placed each session in a directory and labelled it accordingly (Suit, BL, Therapy, Mid, Post, and Maint). Typically developing children recorded a single session each, so the UXTD directories are labelled with speaker identifiers only. Within each session, we sorted the utterances by the date-time of recording and indexed them from 001. We then appended the prompt type identifier (A-F) to the index. For example, if the 5th utterance in a session is a sentence, the tuple of files associated with the utterances are 005C.txt, 005C.wav, 005C.ult, and 005C.param.

3.3. Train, Development, and Test Splits

We split the UXTD dataset into training, development, and testing subsets balancing by gender and age. Training contains 40 children (18 male, 22 female), development 6 children (3 male, 3 female), and testing 12 children (6 male, 6 female). The mean and standard deviation of the children’s age in each subset is: 9y 5m ± 1y 10m, 9y 1m ± 1y 9m, and 8y 12m ± 1y 10m. We omit a split for UX SSD and UPX due to the small number of participants in each SSD subcategory. However, we urge users of this data to report the ID of the participants and the name of sessions used for training and testing.

4. Data Annotation

In addition to the data described in the previous section, we release a set of annotations, including pronunciation dictionaries for each of the datasets, audio transcriptions for UXTD, SLT annotations, automatic speaker labelling and automatic phone alignments, all of which can aid modelling.

Pronunciation dictionaries: We prepared a pronunciation dictionary for each of the three datasets. We did this by listing the words that appear in the prompts, looking them up in a standard lexicon, and copying their phonetic transcription. We used a Scottish accent variant of the Comiblex lexicon to match the accent in the data [29] [30]. For out-of-vocabulary words, such as the non-words of type B prompts, an annotator with training in phonetics transcribed their expected pronunciation. The vocabulary size is 296, 1048, and 1437 for the UXTD, UX SSD, and UPX datasets respectively.

SLT labelling: The SLTs annotated a small portion of the data for the intervention studies the data was originally collected for [24] [25] [26]. The annotations include boundaries of words and phones of interest, and tongue contours manually fitted to the mid-frame of phones of interest in the ultrasound video. The number of utterances with at least one label are 3900, 152, and 3919 in the UXTD, UX SSD, and UPX datasets, respectively. We release these annotations as Praat’s TextGrid files [31] and follow the same naming convention described in Section 3.2.

Audio Transcriptions: Because the audio recording is not a direct match to the prompt, we provide a small subset of transcribed utterances, namely utterances of types A and B for all speakers in the UXTD dataset. A single annotator listened to the audio and transcribed the child’s speech. SLT intervention is loosely transcribed as [SLT:token], where token takes the form of *spn* (spoken sound) to denote generic SLT speech, or the form of a word if that word occurs in the prompt, for example [SLT:helicopter]. It is less obvious how to transcribe disordered speech, we therefore do not provide transcriptions for the UX SSD and UPX datasets.

Speaker labelling: In order to attribute different parts of the audio to different speakers, and to quantify the hours of speech, we trained a model that discriminates between SLT and child speech. We used the transcriptions of the UXTD dataset as training data by reducing words to *child* and *SLT* tokens, corresponding to turn-taking sequences between therapist and participant. Using Kaldi’s [23] standard monophone recipe, we modelled these tokens with 5-state ergodic HMMs [33]. Silences were modelled with 5 state left-to-right skip HMMs. As a post-processing step, we merged identical labels that were separated by a silence with less than 100ms. A second pass of the data then removed labels with duration less than 50ms.

To measure the accuracy of this method, we used the force-aligned transcriptions of the UXTD’s test set as a ground truth. We estimated error using *pyannote.metrics* [34], computing error in terms of seconds. We observed an Identification Error Rate of 4.6%, and precision and recall of 0.969 and 0.979, respectively. We decoded the three datasets with this method, which forms the basis for the data reported in Table 3.

Phone labelling: Automatically identifying phones in a child’s speech would significantly reduce the workload for an SLT. As an initial solution, we applied standard phone alignment to the data. To obtain additional training data, we pooled the training subset of UXTD and the PF-STAR corpus [35], and trained a phone alignment model following the PF-STAR baseline recipe presented in [15].

Using our pronunciation dictionaries, we substituted the words in the prompts and audio transcriptions with their phonetic transcriptions for utterances of types A, B, and C. We then aligned the waveforms to the transcriptions in UXTD, and to
the code repository and invite users to contribute their own code.

We distribute a code repository containing a set of tools to inter-
pret, transform and visualise the data, in addition to the recipes
we release the Kaldi recipes which we used to transform and visualise the
data as the corresponding ultrasound may be useful for other
tasks, such as tongue contour extraction. Figure 1 illustrates this
process. Another tool produces a sequence of images or a video
from a raw ultrasound file.

Recipes: We release the Kaldi recipes which we used to
train the speaker and phone labelling models.

7. License and Distribution
We distribute UltraSuite under Attribution-NonCommercial 4.0
Generic (CC BY-NC 4.0) and distribute the companion code un-
der Apache License v.2. Both can be obtained from the project
website: [http://www.ultrax-speech.org/ultrasuite](http://www.ultrax-speech.org/ultrasuite)

8. Conclusions and Future Work
We have introduced a new repository of ultrasound and acous-
tic data which we have collected from child speech therapy
sessions. We have described the process of data collection,
preparation and standardisation, along with a suggested train-
test split. We have described tools to transform and visualise the
data, and annotations including pronunciation dictionaries, au-
dio transcriptions, SLT annotations, automatic speaker labelling
and automatic phone alignments.

We will continue to grow the repository by adding more
data and tools. We are in the process of collecting further data
from 120 children with SSD in the Ultra2020 project follow-
ing the protocol described in [35]. In addition, we intend to add
other available data to our repository, including adult data and
alternative forms of articulatory imaging techniques (e.g., MRI
of vocal tracts), all of which can be used in data augmentation
methods [17] [15] [10]. We encourage other researchers to con-
tribute by submitting their data for us to standardise and add to
this repository.

9. Acknowledgements
Supported by: EPSRC Healthcare Partnerships Programme,
grants number EP/P027696/1 (Ultrax) and EP/P02338X/1 (Uli-
trax2020), and NHS Scotland CSO, grant number ETM/402 (Uli-
traPhonix). We thank Steve Renals for continued guidance and
support, Anna Womack for help collecting the UltraPhonix data
and Steve Cowen for technical support. We thank the children
for participating and their guardians for providing consent.

### Table 3: Hours of speech and silence rounded to two decimal
places, estimated using the speaker labelling method described
in Section 4, with percentages given in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>UXTD</th>
<th>UXSSD</th>
<th>UPX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child speech</td>
<td>2.24 (28.39%)</td>
<td>3.66 (34.45%)</td>
<td>7.27 (38.70%)</td>
</tr>
<tr>
<td>SLT speech</td>
<td>1.24 (15.74%)</td>
<td>1.81 (16.99%)</td>
<td>1.92 (10.23%)</td>
</tr>
<tr>
<td>Total speech</td>
<td>3.47 (44.12%)</td>
<td>5.47 (51.43%)</td>
<td>9.19 (48.93%)</td>
</tr>
<tr>
<td>Initial silence</td>
<td>1.41 (17.96%)</td>
<td>0.91 (8.55%)</td>
<td>0.78 (4.17%)</td>
</tr>
<tr>
<td>Medial silence</td>
<td>1.99 (25.30%)</td>
<td>3.48 (32.69%)</td>
<td>7.11 (37.83%)</td>
</tr>
<tr>
<td>Final silence</td>
<td>0.99 (12.61%)</td>
<td>0.78 (7.32%)</td>
<td>1.70 (9.07%)</td>
</tr>
<tr>
<td>Total silence</td>
<td>4.40 (55.88%)</td>
<td>5.16 (48.57%)</td>
<td>9.59 (51.07%)</td>
</tr>
<tr>
<td>Total audio</td>
<td>7.87</td>
<td>10.63</td>
<td>18.78</td>
</tr>
</tbody>
</table>

Figure 1: An ultrasound image showing the midsagittal view of
a child’s tongue. We store the raw ultrasound reflection data
efficiently as a matrix (left), but provide a tool to transform it to
real world proportions (right).
10. References


