TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos

Citation for published version:

Digital Object Identifier (DOI):
10.1109/SLT48900.2021.9383619

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2021 IEEE Spoken Language Technology Workshop (SLT)

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.
TAL: A SYNCHRONISED MULTI-SPEAKER CORPUS OF ULTRASOUND TONGUE IMAGING, AUDIO, AND LIP VIDEOS

Manuel Sam Ribeiro¹, Jennifer Sanger¹, Jing-Xuan Zhang¹,², Aciel Eshky¹, Alan Wrench³,⁴, Korin Richmond³, Steve Renals¹

¹The Centre for Speech Technology Research, University of Edinburgh, UK,
²University of Science and Technology of China, P.R.China,
³Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, UK,
⁴Articulate Instruments Ltd., UK

{sam.ribeiro, korin.richmond, s.renals}@ed.ac.uk

ABSTRACT

We present the Tongue and Lips corpus (TaL), a multi-speaker corpus of audio, ultrasound tongue imaging, and lip videos. TaL consists of two parts: TaL1 is a set of six recording sessions of one professional voice talent, a male native speaker of English; TaL80 is a set of recording sessions of 81 native speakers of English without voice talent experience. Overall, the corpus contains 24 hours of parallel ultrasound, video, and audio data, of which approximately 13.5 hours are speech. This paper describes the corpus and presents benchmark results for the tasks of speech recognition, speech synthesis (articulatory-to-acoustic mapping), and automatic synchronisation of ultrasound to audio. The TaL corpus is publicly available under the CC BY-NC 4.0 license.

Index Terms— Ultrasound tongue imaging, video lip imaging, silent speech, articulography, corpora

1. INTRODUCTION

Measuring the position of the articulators during the speech production process is relevant to disciplines such as linguistics, speech processing, speech pathology, and anatomy. Articulatory movement can be captured using magnetic articulography techniques, such as electromagnetic articulography (EMA) or permanent-magnetic articulography (PMA); palatography techniques, such as electropalatography or optopalatography; or imaging techniques, such as video imaging, magnetic resonance imaging (MRI) or ultrasound tongue imaging (UTI).

Video is the cheapest and most convenient imaging method to acquire articulatory data and is useful for studies focusing on the dynamics of those articulators. Its main limitation is its inability to capture anything beyond extracranial articulators, such as the lips and jaw. To provide complementary data, intraoral articulators can be monitored using medical imaging techniques such as MRI [2] and UTI [3]. Although MRI captures high-quality images, it suffers from a variety of disadvantages [4]: it is expensive and not easily accessible, and suffers from loud background noise, a supine recording position, and low temporal resolution. Ultrasound tongue imaging, on the other hand, is portable, non-invasive, clinically safe, and relatively cheap. UTI uses diagnostic ultrasound operating in B-mode to visualize the tongue surface during speech production at high frame rates. There are, however, some challenges associated with ultrasound images of the tongue [5]. They may be noisy and image resolution may not be ideal, as they are susceptible to unrelated high-contrast edges, speckle noise, or interruptions of the tongue surface. Furthermore, image quality may be affected by speaker characteristics such as age or physiology, or by probe placement.

Ultrasound tongue imaging has been used in various applications, including speech therapy [6–7, 8], language learning [9] [10], phonetics studies [11], and the development of silent speech interfaces [11]. Previous work in the context of speech therapy used ultrasound data to develop tongue contour extractors [12], animate a tongue model [13], and automatically synchronise and process speech therapy recordings [14] [15]. Additionally, speech recognition [16, 17] and speech synthesis [18] [19] from ultrasound images have been used in silent speech interfaces to restore spoken communication for users with voice impairments or to allow silent communication in situations where audible speech is undesirable. Ultrasound data has also been used to develop models for articulatory-to-acoustic or acoustic-to-articulatory mapping [20, 21]. Lip videos have been used for research on multimodal speech perception [22], audio-visual synchronisation [23], automatic lip reading [24], audio-visual speech recognition [25], and speech reconstruction from silent videos [26].

There are relatively few publicly available corpora of ultrasound tongue images. The Silent Speech Challenge dataset [27] consists of approximately 2500 utterances of ultrasound and video images from a single native English speaker. The UltraSuite repository [28] contains ultrasound and audio data of 58 typically developing children and 28 children with speech sound disorders in the context of speech therapy. Because video imaging is easier to acquire, there is a much larger selection of audio-visual datasets of lip videos (see [24] [29] for recent surveys). Earlier corpora were concerned with restricted-vocabulary tasks, such as the recognition of the alphabet [30], isolated digits [31], or sentences with limited vocabulary [32]. Recent work, however, is based on more comprehensive audio-visual datasets containing several hundred hours of speech and thousands of speakers [33, 34].

This paper presents the Tongue and Lips Corpus (TaL), which contains synchronised imaging data of extraoral (lips) and intraoral...
(tongue) articulators from 82 native speakers of English. Section 2 describes the data collection and data post-processing steps of the TaL corpus. Section 3 introduces a set of benchmark experiments for automatic speech recognition, speech synthesis, and automatic synchronisation of ultrasound and audio. We conclude in Section 4 with an analysis of speaker performance across tasks and systems.

2. THE TAL CORPUS

The TaL corpus (Table 1) comprises TaL1, a set of six recording sessions of one professional voice talent, a male native speaker of English; and TaL80, a set of recording sessions of 81 native speakers of English, without voice talent experience. All recording sessions took place between October 2019 and March 2020. Sessions with the experienced voice talent were approximately 120 minutes in duration, and sessions with the remaining speakers were approximately 80 minutes.

<table>
<thead>
<tr>
<th>TaL1</th>
<th>TaL80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of speakers</td>
<td>1</td>
</tr>
<tr>
<td>Gender (male/female)</td>
<td>1/0</td>
</tr>
<tr>
<td>Accent (eng/sco/other)</td>
<td>1/0/0</td>
</tr>
<tr>
<td>Age (mean/std)</td>
<td>61/-</td>
</tr>
<tr>
<td>Utterances per session (mean/std)</td>
<td>263/14</td>
</tr>
<tr>
<td>Total number of utterances</td>
<td>1582</td>
</tr>
</tbody>
</table>

Table 1. TaL corpus at a glance. Gender, accent, and age are self-reported. Accent is categorised here in terms of English (eng), Scottish (sco), or other.

Prompts Type Shared Count

| Swallow | - | Yes | 1 |
| Calibration | Read | Yes | 2 |
| Rainbow, Accent | Read | Yes | 24 |
| Calibration | Silent | Yes | 2 |
| Rainbow (partial) | Silent | Yes | 6 |
| Harvard | Silent | No | 9 |
| Harvard | Read | No | 9 |
| Spontaneous | - | No | 1 |
| TIMIT, VCTK, Librispeech | Read | No | - |

Table 2. Recording protocol for TaL80. “Type” indicates speech type (read or silent speech). “Shared” indicates that the prompts were read by all speakers in TaL80. The last row does not have a defined number of prompts, as participants read as many as possible until the end of the session.

Participants were asked to swallow and to read two calibration sentences at the beginning and end of each recording session, and before and after any breaks. The protocol includes a mixture of prompts read by all speakers and prompts read by each speaker. Some prompts were read silently, audibly, or both. Additionally, each participant was asked to speak unprompted for 30-60 seconds. A list of suggested conversation topics was available, although participants were allowed to choose any topic of their preference. Each TaL1 session follows a similar protocol, with the addition of whispered speech utterances.

Using the AAA software, we synchronised the video and ultrasound data to the audio stream. During the synchronisation process, the video stream was deinterlaced. We then exported all recorded data from the AAA software. We normalised waveform levels using the tools available in ffmpeg. Using ffmpeg, the video data was scaled to 240 × 320 pixels and encoded with libx264 using yuv420p pixel format and a bit rate of 1 Mbps. Due to a configuration error, the synchronisation signal was corrupted for the first recording session of TaL1 (day1). For this reason, video synchronisation is not available for this session.

For each speaker, utterances were sorted by recording date-time and indexed from 001. Each filename contains an additional tag, denoting its prompt type. The tag swa is used for swallows, cal for calibration utterances, spo for spontaneous speech, sil for silent speech, and aud for audible read speech. The marker x was used with audible and silent tags to indicate cross-prompts (repeated by all speakers). The absence of this marker indicates that the prompt is unique to the current speaker. The TaL1 data contains the additional whi tag to indicate whispered speech. Cross-prompt tags in this data indicate prompts repeated across sessions. Furthermore, each utterance consists of five core data types identified by file extension:

1. The prompt file is identified by extension .txt and it includes the prompt text read by the speaker, and the date and time of the recording.
2. The waveform, identified by the file extension .wav, is a single-channel RIFF wave file, sampled at 48 KHz with a bit-depth of 16-bit.
3. The synchronization signal is stored identically to the speech waveform, but is identified by the file extension .sync. This waveform contains the audio pulses used to synchronise the video and the ultrasound stream.
Table 3. Amount of data for the TaL corpus by prompt type, computed over parallel data streams (audio, video, ultrasound). The speech columns estimate data after Voice Activity Detection. All estimates are given in minutes, except for the last row, which is given in hours. Shared data (denoted by x in the tag identifier) means that the same set of prompts are read by all speakers. For TaL1, x is applied across recording sessions.

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>Tag</th>
<th>TaL1 speech</th>
<th>TaL1 total</th>
<th>TaL80 speech</th>
<th>TaL80 total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>aud</td>
<td>53.26</td>
<td>88.93</td>
<td>501.83</td>
<td>829.45</td>
</tr>
<tr>
<td>Silent</td>
<td>sil</td>
<td>0.00</td>
<td>6.31</td>
<td>0.00</td>
<td>62.47</td>
</tr>
<tr>
<td>Whispered</td>
<td>whi</td>
<td>2.98</td>
<td>6.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Read (shared)</td>
<td>xaud</td>
<td>4.04</td>
<td>5.70</td>
<td>162.18</td>
<td>218.13</td>
</tr>
<tr>
<td>Silent (shared)</td>
<td>xsil</td>
<td>0.00</td>
<td>6.37</td>
<td>0.00</td>
<td>78.19</td>
</tr>
<tr>
<td>Whispered (shared)</td>
<td>xwhi</td>
<td>2.62</td>
<td>3.83</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>spo</td>
<td>3.68</td>
<td>4.74</td>
<td>46.50</td>
<td>56.24</td>
</tr>
<tr>
<td>Calibration</td>
<td>cal</td>
<td>2.98</td>
<td>4.60</td>
<td>32.60</td>
<td>48.90</td>
</tr>
</tbody>
</table>

Total: 1.16 hrs 2.14 hrs 12.39 hrs 21.90 hrs

4. The ultrasound data is stored across two files. Raw ultrasound data is identified by the file extension .ult, while the extension .param is a text file containing ultrasound metadata (e.g., frames per second, synchronization offset, etc.).

5. The video data is identified by the file extension .mp4, which embeds its metadata.

We transcribed all spontaneous speech utterances and included the text transcription in the corresponding prompt file. Because spontaneous speech utterances can be long in duration (up to 60 seconds), we manually annotated the boundaries of shorter time segments (typically 5-10 seconds). This segmentation was added as an additional data type for each speaker, identified by the file extension .lab. This additional data type, available only for spontaneous speech utterances, is a text file with the start and end time in seconds for each time segment and their respective transcriptions.

The AAA software receives data from each modality separately. The ultrasound stream begins recording after the video and audio and normally stops recording after. We release all data streams without trimming them, as they may be useful for a variety of tasks which are not dependent on parallel data. Identifying parallel data should be trivial, given the available metadata for each data stream.

For a summary of the TaL corpus, we refer the reader back to Figure 1. Age distribution for TaL80.

4. The ultrasound data is stored across two files. Raw ultrasound data is identified by the file extension .ult, while the extension .param is a text file containing ultrasound metadata (e.g., frames per second, synchronization offset, etc.).

5. The video data is identified by the file extension .mp4, which embeds its metadata.

We transcribed all spontaneous speech utterances and included the text transcription in the corresponding prompt file. Because spontaneous speech utterances can be long in duration (up to 60 seconds), we manually annotated the boundaries of shorter time segments (typically 5-10 seconds). This segmentation was added as an additional data type for each speaker, identified by the file extension .lab. This additional data type, available only for spontaneous speech utterances, is a text file with the start and end time in seconds for each time segment and their respective transcriptions.

The AAA software receives data from each modality separately. The ultrasound stream begins recording after the video and audio and normally stops recording after. We release all data streams without trimming them, as they may be useful for a variety of tasks which are not dependent on parallel data. Identifying parallel data should be trivial, given the available metadata for each data stream.

For a summary of the TaL corpus, we refer the reader back to Table 1. Table 3 includes more a detailed distribution of the amount of data by prompt. Overall, we have collected roughly 24 hours of parallel ultrasound, video, and audio, of which approximately 13.5 hours are speech. This was collected for a total of 82 speakers across 87 recording sessions. TaL1 contains roughly 2 hours of data from a single experienced voice talent across six recording sessions. TaL80 contains almost 22 hours of data from 81 native speakers of English. Figure 1 shows the distribution of ages for TaL80. There is a slight bias towards younger speakers, as participants were recruited primarily via University channels. Figure 2 shows sample frames from twelve speakers available in TaL80.

### 3. EXPERIMENTS

In this section, we present a set of benchmark experiments for the TaL corpus, focusing on the available modalities. Section 3.1 presents initial results for automatic speech recognition. Section 3.2 provides results for the articulatory-to-acoustic mapping task, and Section 3.3 presents results for the automatic synchronisation of ultrasound and audio.

#### 3.1. Speech recognition

This set of experiments provides results for speech recognition using the three modalities available in the TaL corpus. We further consider these results across two training scenarios: speaker dependent and speaker independent. For this task, we use only read speech utterances (identified by aud), and discard silent, spontaneous, whispered, and other shared prompts. We expect speech styles to have different properties in terms of tongue and lip movement, therefore we focus only on read speech.

For the speaker dependent scenario, we conduct our experiments with the TaL1 dataset. We discard day1, for which video synchronisation is not available. The remaining sessions, days 2-6, are evaluated using cross-validation, leaving one recording session out for testing. We aggregate results by computing WER on the total number of errors and words rather than averaging WER across the 5 folds. For the speaker independent scenario, we train models on speakers 11-81 from the TaL80 dataset. We discard from the training set all prompts that co-occur with the TaL1 dataset. We save the first 10 speakers for testing, as these have the highest prompt overlap with TaL1. Days 2-6 from TaL1 are used as test set, which is identical to the cross-validation setup for the speaker dependent system.

**Feature extraction.** During data preparation, calibration utterances are used to identify breaks in the recording session or adjustments to the equipment. This may indicate different image configurations, so we define a “speaker” to be any set of utterances without any breaks. This arrangement allows us to do speaker adaptive training on a speaker dependent scenario. We extract features over parallel data streams and we trim initial and end silences in the ultrasound and video modalities after Voice Activity Detection on the audio stream. For the audio stream, we extract 13 dimensional Mel Frequency Cepstral Coefficients (MFCCs) and we append their respective first and second derivatives. For the visual stream, we apply the 2D Discrete Cosine Transform (DCT) to each ultrasound or video frame. The upper left submatrix is flattened and used as input. Video frames are resized to 120 × 160 and ultrasound frames are resized to 64 × 420. Both streams use a submatrix of 12 × 12 coefficients (144 coefficients).
Table 4. Word Error Rate (WER) for speaker dependent and speaker independent speech recognition on TaL1 and TaL80.

<table>
<thead>
<tr>
<th>Data streams</th>
<th>Speaker Dependent</th>
<th>Speaker Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TaL1</td>
<td>TaL1</td>
</tr>
<tr>
<td>Audio</td>
<td>3.42%</td>
<td>4.06%</td>
</tr>
<tr>
<td>Tongue</td>
<td>38.41%</td>
<td>75.34%</td>
</tr>
<tr>
<td>Lips</td>
<td>51.57%</td>
<td>84.70%</td>
</tr>
<tr>
<td>Tongue+Lips</td>
<td>25.09%</td>
<td>49.12%</td>
</tr>
</tbody>
</table>

Model training. We train the “acoustic” models with the Kaldi speech recognition toolkit [45]. For these experiments, we initialise models from a flat start using the corresponding features. After monophone and triphone training, input features are processed with Linear Discriminant Analysis (LDA) and a Maximum Likelihood Linear Transform (MLLT). This is followed by Speaker Adaptive Training (SAT) with feature-space MLLR (fMLLR) [46]. The alignment and HMM models from this stage are then used to train a time-delay neural network (TDNN, [47]) following Kaldi’s nnet3 recipe. We decode with a simple in-domain bigram language model trained on all prompts from the sources described in section 2.

Results and future work. Results are presented in Table 4. As expected, recognition results from the audio stream outperform those of the articulatory modalities by a large margin. Although we have made no attempts to design more complex systems for silent speech recognition, the high WERs illustrate the difficulty of the problem. This is particularly noticeable when using data from multiple speakers in a speaker independent system. In all cases, we observe that the visual data streams complement each other well. The difference in the TaL1 results across the two scenarios shows that speaker independent silent speech recognition is a challenging problem. Additionally, we observe a large range in the speaker-wise WERs in the TaL80 test set. Using the Tongue+Lips system, the best speaker achieves a WER of 32.3% and the worst speaker a WER of 98.3%. Further work can focus on identifying speakers with poor image quality and developing methods to improve their results. There are a number of possibilities that could improve upon these results. For example, we made no attempt to use the audio stream to initialise the silent models. Instead of a flat start, we could bootstrap monophone HMMs with audio alignment. Earlier work showed that substantial improvements can be observed with more complex feature extraction [17, 48]. The results on the TaL80 corpus are speaker independent, but they do make some use of data from target speakers when estimating fMLLR transforms. We observe that using fMLLR can lead to good improvements. Other forms of speaker adaptation are likely to be useful. The multi-speaker systems and fine-tuning results described in the next section further support this idea.

3.2. Speech synthesis

This set of experiments presents results for an articulatory-to-acoustic conversion model. As in Section 3.1, we show results in terms of the three modalities available in the TaL corpus. Instead of a speaker-independent scenario, we focus here on speaker-dependent and multi-speaker models for speech recovery from silent modalities. As before, the speaker-dependent model uses days 2-6 from TaL1. In the multi-speaker condition, we use data from 75 speakers in TaL80 and we hold out the rest for additional experiments. For training data, we use all the aud utterances available for each speaker, of which 10 utterances are randomly held-out for validation. We use the 24 xaud utterances from each speaker for testing.

Feature extraction. The video stream is resampled to match the frame rate of the ultrasound stream using ffmpeg. Each video frame is then resized to \( 72 \times 136 \), randomly flipped horizontally, then cropped to \( 64 \times 128 \) pixels. Each ultrasound frame is resized to \( 64 \times 128 \). For acoustic features, we use the STRAIGHT vocoder [49] to extract 41-dimensional Mel-cepstral coefficients (MCCs) and
on each speaker. We used the Adam optimizer with a learning rate pre-trained on all 75 training speakers and then fine-tuned separately. The representations from the two encoders are concatenated conditioned on the encoder outputs and speaker x-vectors. When dimensional vectors. The decoder then predicts acoustic features.

The encoder transforms the tongue or lip frames into 512-dimensional vectors. The decoder then predicts acoustic features normalized to have zero mean and unit variance for each speaker. In the multi-speaker model, we additionally use a speaker representation in the form of x-vectors [50], extracted with the Kaldi toolkit. For wave-form reconstruction, we use the STRAIGHT vocoder.

Model training. We use an encoder-decoder architecture (Table 5). The encoder transforms the tongue or lip frames into 512-dimensional vectors. The decoder then predicts acoustic features conditioned on the encoder outputs and speaker x-vectors. When both tongue and lip videos are used as inputs, we adopt separate encoders. The representations from the two encoders are concatenated and sent to the decoder. For multi-speaker training, the model is first pre-trained on all 75 training speakers and then fine-tuned separately on each speaker. We used the Adam optimizer with learning rates of $10^{-3}$ and $10^{-4}$ during pre-training and fine-tuning, respectively.

Results and future work. Results are presented in Table 6. To evaluate the performance of the model, we report results in terms of mel-cepstral distortion (MCD). As a proxy measure of intelligibility, we decode the synthesized samples with an open-source ASR model based on ESPNet [51] and we report results in terms of WER. We opt to use this model rather than the audio model described in section 5.1, as it is not trained on audio data from the target speakers. Decoding the corresponding natural speech with this model achieves a WER of 70.6% for TaL1 and TaL80, respectively. We observe from Table 6 that the performance of the model using different modalities is similar to the results reported in Section 5.1. Using ultrasound tongue images leads to better results than video images of the lips. The best models use both the ultrasound and video data. As with speech recognition, the average performance of the multi-speaker system is worse than that of a speaker-dependent system. We further provide results for fine-tuning the multi-speaker model on the TaL1 data. This is denoted by TaL80+TaL1 in Table 6. Future work will evaluate the intelligibility of synthetic speech with a perceptual test. To this end, a neural vocoder can be used to generate high quality waveforms. More powerful encoders, such as ResNet [52], can be used to process the video inputs. As observed in section 5.1, results vary substantially across speakers. Focusing on identifying speakers whose converted speech has low intelligibility and optimizing for their performance could be an interesting future research direction.

### Table 5. Model structure for the articulatory-to-acoustic model.

| Encoder          | 3DCNN, $k \times 5 \times 5$, $s (1, 2, 2)$, $c \rightarrow$ 32 $\rightarrow$ BN-ReLU-Dropout $0.2 \rightarrow$ Maxpooling, $k \times 2 \times 2$, $s (1, 2, 2)$ $\rightarrow$ 3DCNN, $k \times 5 \times 5$, $s (1, 2, 2)$, $c \rightarrow$ 64 $\rightarrow$ BN-ReLU-Dropout $0.2 \rightarrow$ Maxpooling, $k \times 2 \times 2$, $s (1, 2, 2)$ $\rightarrow$ 3DCNN, $k \times 3 \times 3$, $s (1, 2, 2)$, $c \rightarrow$ 128 $\rightarrow$ BN-ReLU-Dropout $0.2 \rightarrow$ Flatten $\rightarrow$ FC-512 units $\rightarrow$ BN-ReLU-Dropout $0.2$
| Decoder          | 2-layer BLSTM, 256 units in each direction $\rightarrow$ FC-43 units

$k$ and $s$ represent kernel size and stride respectively. $c$ represents the number of output channels. The dimension order is $T$ (time), $H$ (height), $W$ (width). BN represents batch normalization. FC represents fully connected layer.

### Table 6. Mel-Cepstral Distortion (MCD) and Word Error Rate (WER) for the articulatory-to-acoustic conversion model. In the dataset column, TaL1 denotes results for a speaker-dependent model, TaL80 for a multi-speaker model, and TaL80+TaL1 indicates the multi-speaker model fine-tuned on all available TaL1 data. For TaL80, we show mean ± standard deviation across all speakers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data streams</th>
<th>MCD (dB)</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaL80</td>
<td>Lips</td>
<td>4.09±0.19</td>
<td>96.0±3.8</td>
</tr>
<tr>
<td></td>
<td>Tongue</td>
<td>3.35±0.16</td>
<td>56.5±11.0</td>
</tr>
<tr>
<td></td>
<td>Tongue+Lips</td>
<td>3.31±0.15</td>
<td>53.5±11.7</td>
</tr>
<tr>
<td>TaL1</td>
<td>Lips</td>
<td>3.43</td>
<td>66.3</td>
</tr>
<tr>
<td></td>
<td>Tongue</td>
<td>2.99</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>Tongue+Lips</td>
<td>2.84</td>
<td>17.2</td>
</tr>
<tr>
<td>TaL80+TaL1</td>
<td>Tongue+Lips</td>
<td>2.72</td>
<td>14.0</td>
</tr>
</tbody>
</table>

### Table 7. Accuracy of UltraSync, when pre-trained on out-of-domain child speech therapy data, and when trained on in-domain TaL data.

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>Tag</th>
<th>TaL1</th>
<th>TaL80</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n accuracy</td>
<td>n accuracy</td>
<td></td>
</tr>
<tr>
<td><strong>Out-of-domain model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>aud</td>
<td>384</td>
<td>70.6%</td>
</tr>
<tr>
<td>Read (shared)</td>
<td>xaud</td>
<td>48</td>
<td>79.2%</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>spo</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Calibration</td>
<td>cal</td>
<td>18</td>
<td>77.8%</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>452</td>
<td>71.9%</td>
</tr>
<tr>
<td><strong>In-domain model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>aud</td>
<td>384</td>
<td>98.4%</td>
</tr>
<tr>
<td>Read (shared)</td>
<td>xaud</td>
<td>48</td>
<td>97.9%</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>spo</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Calibration</td>
<td>cal</td>
<td>18</td>
<td>100%</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>452</td>
<td>98.5%</td>
</tr>
</tbody>
</table>

3.3. Automatic Synchronisation

These experiments present results for the automatic synchronisation of ultrasound and audio. The TaL corpus was synchronised at recording time using a hardware synchronisation mechanism which gives the offset between the two signals in milliseconds. We investigate how well we can predict these offsets by exploiting correlations between the two modalities using the UltraSync architecture [14].

We conducted two experiments using in and out-of-domain data. For the first experiment, we use the UltraSync model from [14] which has been trained on out-of-domain child speech therapy data. In the second experiment, we train the UltraSync model on in-domain TaL data. We reserve days 2, 3, and 4 from TaL1 for training, day 5 for validation, and day 1 and 6 for testing. We also reserve speakers 1-49 from TaL80 for training, 50-65 for validation, and 66-81 for testing. We pool the training data and train a single in-domain model from scratch, following the same architecture and training procedures as [14]. We report all of our results on the same test subset for comparability.

Data preprocessing. We preprocess the data to match the Ultra-
Sync input size by resampling the audio to 22.05 KHz using `scipy interpolate`, resampling the ultrasound to 24fps using `skimage transform`, and resizing the ultrasound frames to $65 \times 138$ pixels.

**Experimental setup.** UltraSync requires us to specify the range of allowable offsets. In a practical scenario, domain knowledge is utilised to select a suitable candidate list [14]. For these experiments, we consider the observed minimum and maximum offsets given by the hardware sync in the TaL corpus and set the step size to 45ms. We then calculate the range of allowable offsets as: $[\min - (\text{step} \times 10), \max + (\text{step} \times 10)]$, which renders 25 candidates in total, roughly equal to the 24 candidates in [14].

**Results and future work.** For each utterance we calculate the discrepancy: $\text{disc} = \text{prediction} - \text{truth}$, where $\text{truth}$ is the hardware offset. A predicted offset is correct if the discrepancy falls within the detectability threshold: $-125 < \text{disc} < +45$ [14]. We report results on utterances with audible content and exclude silent and whispered speech in Table 7. The overall out-of-domain model accuracy is 72.2% while the in-domain accuracy is 97.7%, compared to 82.9% reported by [14]. We attribute the increase in performance to the higher quality of our data, the absence of multiple speakers in our recordings, and the presence of linguistic variety in our utterances compared to child speech therapy data [14]. Future work will explore synchronising the two visual modalities or all three modalities.

### 4. DISCUSSION

Experimental results show that results can vary depending on tasks or articulatory modalities. In this section, we perform an analysis of speaker performance. We prepare a common evaluation set of all xaud utterances available in TaL80 (24 per speaker), which were not seen during training. We collect results separately for each speaker using the corresponding speech recognition and articulatory-to-acoustic mapping systems. We use WER from the TaL80 speaker-independent system in Section 3.3 and MCD from the TaL80 multi-speaker system of Section 4.4. Given these speaker-wise WER/MCD scores, we compute Pearson's product-moment correlation across all system pairs (Figure 3). We observe that speaker performance is generally more correlated for intra-task systems than inter-task systems, irrespective of modality used. For example, considering articulatory-to-acoustic mapping (MCD in Figure 3), speaker performance correlates well when using different modalities. That is, speaker-wise MCD results for the systems using tongue-only and lips-only have medium correlation $(r = 0.51)$. But when comparing speaker performance across tasks using the same modality, correlation tends to be lower, either with tongue inputs $(r = 0.34)$ or lip inputs $(r = 0.21)$. We also observe a very high correlation for the articulatory-to-acoustic mapping system using tongue and tongue+lip data $(r = 0.98)$. This suggests that the system may not be making the most use of the video data. This could explain the limited improvements of the tongue+lip system over the tongue system in Section 3.2 when compared with the corresponding system in Section 4.4.

Additionally, we compare speaker performance on systems using only one of the available modalities. To find a comparable score across both tasks, each speaker score (WER or MCD) is replaced by their rank order and then averaged across the two tasks. Figure 4 shows speaker mean ranks for systems using only tongue images or lip images. Although there is low inter-task correlation, Figure 4 identifies speakers that underperform on both modalities and tasks (e.g. $09fe$, $16fe$, $27fs$, $12me$, and $60ms$). Similarly, some speakers perform well across the two data streams and tasks (e.g. $70ms$, $48ms$, and $22me$). On the other hand, speaker $63me$, for example, achieves good score when using ultrasound data, but scores poorly when using only lip data. This type of analysis can be used in the future to identify speakers with good-quality imaging data.

### 5. CONCLUSION

We presented the Tongue and Lips corpus (TaL) and a set of benchmark experiments for speech recognition, articulatory-to-acoustic mapping, and automatic synchronisation of ultrasound and audio. These experiments illustrate the challenge of developing robust systems that process articulatory imaging data from multiple speakers. TaL is released under the Creative Commons Attribution-NonCommercial 4.0 Generic (CC BY-NC 4.0) licence and is distributed via the UltraSuite Repository: https://www.ultrax-speech.org/ultrasuite

![Fig. 3. Pearson’s product-moment correlation of speaker-wise results across system pairs. Each cell in this figure indicates the correlation of speaker performance for two systems. WER denotes results from speech recognition systems and MCD denotes results for articulatory-to-acoustic mapping systems.](image-url)

![Fig. 4. Speaker mean rank for speech recognition and articulatory-to-acoustic mapping using Tongue-only or Lip-only data for TaL80. Speaker identifiers denote speaker number, gender, and accent.](image-url)
6. REFERENCES


