Automatic audiovisual synchronisation for ultrasound tongue imaging

Aciel Eshky[a,*], Joanne Cleland[b], Manuel Sam Ribeiro[a], Eleanor Sugden[b], Korin Richmond[a], Steve Renals[a]

[a]*The Centre for Speech Technology Research, University of Edinburgh, UK*
[b]*Psychological Sciences and Health, University of Strathclyde, UK*

Abstract

Ultrasound tongue imaging is used to visualise the intra-oral articulators during speech production. It is utilised in a range of applications, including speech and language therapy and phonetics research. Ultrasound and speech audio are recorded simultaneously, and in order to correctly use this data, the two modalities should be correctly synchronised. Synchronisation is achieved using specialised hardware at recording time, but this approach can fail in practice resulting in data of limited usability. In this paper, we address the problem of automatically synchronising ultrasound and audio after data collection. We first investigate the tolerance of expert ultrasound users to synchronisation errors in order to find the thresholds for error detection. We use these thresholds to define accuracy scoring boundaries for evaluating our system. We then describe our approach for automatic synchronisation, which is driven by a self-supervised neural network, exploiting the correlation between the two signals to synchronise them. We train our model on data from multiple domains with different speaker characteristics, different equipment, and different recording environments, and achieve an accuracy >92.4% on held-out in-domain data. Finally, we introduce a novel resource, the Cleft dataset, which we gathered with a new clinical subgroup and for which hardware synchronisation proved unreliable. We apply our model to this out-of-domain data, and evaluate its performance subjectively with expert users. Results show that users prefer our model's output over the original hardware output 79.3% of the time. Our results demonstrate the strength of our approach and its ability to generalise to data from new domains.

Keywords: Automatic audiovisual synchronisation, synchronisation error tolerance, ultrasound tongue imaging

1. Introduction

Ultrasound tongue imaging visualises the shape, position, and movement of the tongue during speech production. It is utilised in a number of applications including speech and language therapy, phonetics research, second language learning, and silent speech interfaces (Cleland et al., 2019; Lawson et al., 2015; Wilson et al., 2006; Hueber et al., 2010). In the majority of applications, ultrasound is acquired simultaneously with audio, and for the data to be correctly processed and analysed, the two modalities should be correctly synchronised. Synchronisation can be achieved at recording time using specialised hardware (Hueber et al., 2008), however, this approach can fail in practice resulting in data of limited usability (Cleland, 2018). Furthermore, synchronisation information is not always available for historical data (Bakst & Lin, 2019). While manual synchronisation is possible, it is time consuming, and particularly challenging in the absence of useful audiovisual cues such as stop closures and bursts. To address the lack of a mitigation strategy for the failure of hardware synchronisation, we previously introduced a method to automatically synchronise ultrasound and audio after data collection (Eshky et al., 2019), and to our knowledge, no work prior to ours attempted this. Our approach used a self-supervised neural network which exploits correlations between the two signals to synchronise them without the need for manual annotation.

In this paper, we expand on our previous work. Our first novel contribution is a detailed investigation of the tolerance for synchronisation error by expert ultrasound users. While the tolerance is known for lip video (ITU-R, 1998), no prior work examines it for ultrasound tongue imaging. This investigation allows us to identify the threshold for detecting synchronisation error, and to define ac-

*Corresponding author
Email addresses: a.eshky@ed.ac.uk (Aciel Eshky), joanne.cleland@strath.ac.uk (Joanne Cleland), Sam.Ribeiro@ed.ac.uk (Manuel Sam Ribeiro), eleanor.sugden@strath.ac.uk (Eleanor Sugden), korin.richmond@ed.ac.uk (Korin Richmond), s.renals@ed.ac.uk (Steve Renals)

curacy scoring boundaries for evaluating synchronisation systems.

Our second contribution builds directly on our previous work in Eshky et al. (2019). We adopt the UltraSync architecture, retraining the model on data from multiple domains with different speaker characteristics, different equipment, and different recording environments to give it the best chance of generalising to data from new domains, and evaluate the model in the first instance on held-out data of the same domain.

Our final contribution is a novel dataset which we recorded from children diagnosed with cleft lip and palate; a clinical subgroup not previously examined in the context of automatic audiovisual synchronisation, or indeed, automatic processing. Hardware synchronisation proved unreliable for the Cleft data, making it a prime application candidate for our model. Because this data was collected with a new clinical subgroup, in a different environment, and using varied ultrasound settings, we are able to use it to test our model's ability to generalise. We apply our model to this out-of-domain data, and evaluate its performance subjectively with expert users. As part of this work, we make the dataset available to the research community in open format.

The paper is organised as follows. In Section 2, we cover related background on ultrasound tongue imaging and audiovisual synchronisation. In Section 3, we describe the ultrasound tongue imaging resources we use for our experiments, and introduce our novel dataset, the Cleft data, which was poorly synchronised at recording time. In Section 4, we describe the perceptual experiment we designed to identify the threshold for detecting synchronisation errors for ultrasound and audio. We use these thresholds to evaluate our system in the section that follows. In Section 5, we describe our automatic synchronisation system, then present automatic evaluation on held-out in-domain data. In Section 6, we apply our approach to the Cleft data, and evaluate the output subjectively in a second perceptual experiment. We summarise our findings in Section 7 and conclude with a discussion in Section 8.

2. Background

To put our work in context, we first present background on ultrasound tongue imaging and its main applications. Then, we transition to audiovisual synchronisation, explaining how it is typically achieved and why it can fail in practice. We discuss user tolerance to synchronisation errors, and present previous research on audiovisual synchronisation, including work on lip video and how it relates to ultrasound.

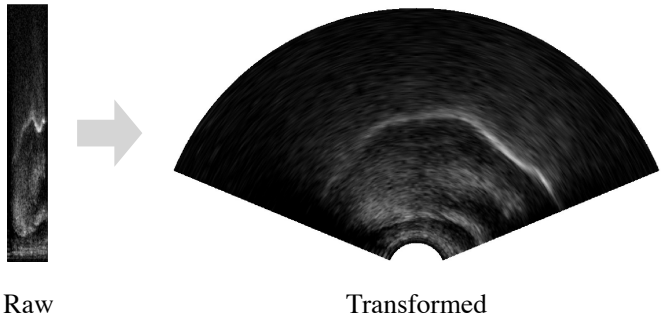

Figure 1: Each ultrasound frame is captured as a matrix of raw reflection data (scan lines × echo returns) and then transformed into real world proportions for visualisation.

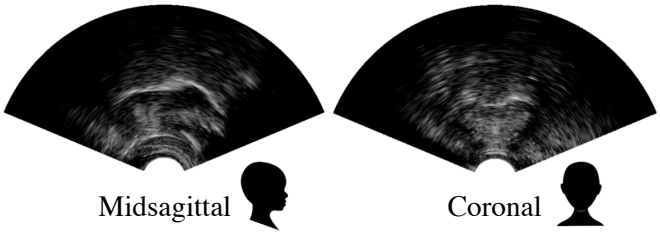

Figure 2: Examples of mid-sagittal and coronal ultrasound tongue images for a child (female, aged 5) with bilateral cleft lip and palate (BCLP), taken from the Cleft dataset (speaker 14). The tip of the tongue is to the right in the mid-sagittal image.

2.1. *Ultrasound tongue imaging*

Ultrasound tongue imaging uses diagnostic ultrasound to visualise the tongue surface. The ultrasound operates in B-mode (brightness mode) in which a linear array of transducers scans a physical surface and returns a matrix of reflection intensities (scan lines × echo returns) for each scan. Ultrasound data can either be stored efficiently as raw reflection data plus the metadata required to transform it into real world proportions for visualisation, or it can be transformed at recording time and stored as videos. Figure 1 shows an example of an ultrasound frame in raw and transformed formats.

To image the tongue, the ultrasound probe is placed under the speaker's chin, capturing either a mid-sagittal or a coronal view of the tongue's surface, depending on the orientation of the probe. Figure 2 shows examples of mid-sagittal and coronal ultrasound tongue images. Ultrasound is clinically safe, non-invasive, portable, and relatively cheap (Gick, 2002; Stone, 2005).

In speech and language therapy, ultrasound tongue imaging can be used to diagnose a range of speech dif-

ficulties, and to provide visual biofeedback in therapy for different types of speech sound disorders, including those arising from a cleft lip or palate (Sugden et al., 2019; Roxburgh et al., 2015; Cleland et al., 2020). During intervention, ultrasound can be used as an objective measure of the patient's progress (Cleland & Scobbie, 2021), or to complement verbal feedback and contribute to positive reinforcement (Roxburgh et al., 2015). Ultrasound also assists annotators in identifying covert articulation errors (Cleland et al., 2017) and has been shown to increase inter-annotator agreement when transcribing the speech of children with cleft lip and palate (Cleland et al., 2020).

Beyond speech therapy, ultrasound is used in phonetics research to compare tongue-shapes for different phones (Davidson, 2006; Lee-Kim et al., 2014; Chen & Lin, 2011; Lawson et al., 2015; Ahn, 2018), or to gain insight into speech production through articulatory-to-acoustic or acoustic-to-articulatory mapping (Hueber et al., 2011; Porras et al., 2019). Ultrasound is also used for practical tasks such as second language learning and acquisition (Wilson et al., 2006; Gick et al., 2008; Mozaffari et al., 2018), or to drive silent speech interfaces, which can be used to restore spoken communication for users with voice impairments (Denby & Stone, 2004; Hueber et al., 2010; Csapó et al., 2017; Ji et al., 2018; Ribeiro et al., 2021b).

To complement this broad range of applications, there is a growing interest in automatically processing and analysing ultrasound, for example, by extracting tongue contours (Fabre et al., 2015; Xu et al., 2016), animating tongue models (Fabre et al., 2017; Chen et al., 2018), classifying speech articulation errors (Ribeiro et al., 2019, 2021a), and most relevant to our work, synchronising it with simultaneously-recorded audio (Eshky et al., 2019).

2.2. *Audiovisual synchronisation*

While ultrasound tongue imaging can be utilised independently, in the majority of applications it is combined with the simultaneously-recorded audio. To be correctly analysed and used, the two modalities should be correctly synchronised. At recording time, specialised hardware captures the relative time difference between the two signals as an offset in milliseconds, and stores it as metadata (Wrench, 2018c,a). Audio leads if the offset is positive, and lags if negative. Applying the offset to an utterance simply involves cropping the leading signal and the end of the trailing signal.

In practice, hardware synchronisation can fail, either as a result of user error, such as incorrectly connecting and operating devices (Cleland, 2018), or as a result of faulty or inferior hardware components, such as low-quality sound cards (Wrench, 2018b), A failure in synchronisation limits the usability of the data (Bakst & Lin, 2019), and while manual synchronisation is possible, it is time-consuming and challenging, especially in the absence of useful audiovisual cues, such as stops and bursts.

User tolerance for synchronisation error depends on the application. Speech and language therapists mainly use recorded ultrasound for playback in intervention sessions to qualitatively evaluate a patient's performance (Cleland et al., 2020), and therefore the synchronisation need only be *perceived* as acceptable by the viewer. In contrast, phoneticians often use acoustic landmarks, such as plosive bursts, to annotate articulatory data, in which case synchronisation should be more precise. Because we work mainly with speech and language therapists, we focus in this paper on the former case.

The majority of research on audiovisual synchronisation focuses on lip videos due to their relevance to broadcasting where synchronisation errors can become objectionable to viewers. In contrast, synchronising audio and ultrasound has received less attention despite its importance. However, because the movement of the articulators (tongue and lips) are correlated (Yehia et al., 1998), we regard prior work on lip synchronisation as relevant. A previous study relying on subjective evaluation found that lip synchronisation errors between -185ms to 90ms are acceptable to viewers, and that the threshold for error detection is -125ms to 45ms (ITU-R, 1998). The study also reported that errors in the range of -95ms to 22.5ms are undetectable to viewers. No such study has been conducted for ultrasound, and therefore the thresholds for detecting synchronisation errors are unknown. In this paper, we address this research gap by examining whether lip thresholds also hold for ultrasound. This investigation allows us to refine our evaluation of automatic synchronisation systems.

Some prior work has been dedicated to automating lip synchronisation. Older approaches investigated the effects of using different representations and feature extraction techniques on finding dimensions of high correlation (Sargin et al., 2007; Bredin & Chollet, 2007; Garau et al., 2010). However, these approaches required extensive feature engineering. More recently, neural networks, which learn features directly from input, have been utilised for the task (Chung & Zisserman, 2016) achieving near-perfect accuracy (99%) on lip synchronisation according to human evaluators. This approach has since been extended to use different methods for creating training samples (Korbar et al., 2018; Chung et al., 2019) and

different model training objectives (Chung et al., 2019).

We previously adopted the original approach from Chung & Zisserman (2016), modifying it for synchronising ultrasound. Our model achieved an accuracy of 82.9% for child speech therapy data (Eshky et al., 2019), and 97.7% for adult speech data (Ribeiro et al., 2021b). In this paper, we build directly on our previous work, training our model on data from multiple domains with different speaker characteristics, different equipment, and different recording environments, and testing our model's ability to generalise to data from a new domain.

3. Data

This section describes the data we use throughout the paper. We first present existing ultrasound datasets which we use for our experiments, then introduce the novel Cleft dataset which we collected with a new clinical subgroup. Hardware synchronisation proved unreliable for the Cleft data, making it a prime candidate to automatically synchronise. We explain the challenges associated with this data and why we class it as a new domain.

Table 1 gives an overview of the data presented in this section. All three resources were recorded in Scotland using the Articulate Assistant Advanced (AAA) software (Articulate Instruments Ltd., 2010), which stores ultrasound efficiently in raw format, augmented with the metadata necessary to transform it into real world proportions for visualisation.

3.1. UltraSuite repository

The first existing resource is the UltraSuite repository (Eshky et al., 2018), which is a collection of three ultrasound and audio datasets gathered from English-speaking children. The data was recorded by research speech and language therapists in a university laboratory. The first dataset is Ultrax Typically Developing (UXTD), collected with 58 typically developing children. The second is Ultrax Speech Sound Disorders (UXSSD), collected with 8 children with speech sound disorders. The third is UltraPhonix (UPX), collected with 20 children with speech sound disorders. The data from UXSSD and UPX was recorded over multiple sessions, including baseline, assessment, therapy, post-therapy, and maintenance.

Ultrasound was recorded using an Ultrasonix SonixRP machine at ≈120fps with a 135° field of view, and the probe was stabilised using a metal headset. Ultrasound frames captured a midsagittal view of the tongue with 63 scan lines × 412 echo returns, and audio was recorded at 22.05 KHz sampling frequency. Audio recordings contained the speech of both the children and therapists. Ultrasound and audio were correctly synchronised at recording time using hardware synchronisation, and this was verified by the researchers who collected the data.

3.2. TaL corpus

The second existing resource is the Tongue and Lips corpus (TaL) (Ribeiro et al., 2021b), which is a collection of ultrasound tongue imaging, lip video, and audio data, recorded with 82 native English-speaking adults. We use the ultrasound and audio for our experiments.

TaL comes in two parts: TaL1 was recorded with a professional voice talent over the course of 6 days, while TaL80 was recorded with 81 speakers with no voice talent experience. Sessions with the voice talent were approximately 120 minutes long, while sessions with the remaining speakers were approximately 80 minutes long. All recordings took place in a hemi-anechoic chamber, resulting in much better audio quality than UltraSuite.

Ultrasound was recorded with a Micro system at ≈80fps with a 92° field of view, and the probe was stabilised using the UltraFit stabilising headset (Spreafico et al., 2018). Ultrasound frames captured a midsagittal view of the tongue with 64 scan lines × 842 echo returns, and audio was recorded at 48 KHz sampling frequency. Ultrasound and audio were correctly synchronised at recording time using hardware synchronisation, and this was verified by the researchers who collected the data.

3.3. Introducing the Cleft dataset

The Cleft dataset is a collection of ultrasound and audio data, gathered from children with cleft lip and palate. The data was recorded by research speech and language therapists in a hospital environment. For this dataset, hardware synchronisation was incorrectly recorded, and was perceived as inadequate by the speech and language therapists who collected the data. In Section 6 we use our system to synchronise the data automatically.

The dataset was originally collected for clinical phonetics research (Cleland et al., 2020) and stored in proprietary format. We processed it, and through this work make it available to the research community in open format. The original data was recorded with 39 English-speaking children, however, only 29 of them gave us consent to share their data (18 male, 11 female). We retain the original speaker IDs for consistency with previous research published on this data (Cleland et al., 2020), but focus in this paper on the 29 speakers whose data we release.

Collection	Dataset	Age	Speech disorder	Environment	Speakers	Stabilisation	Ultrasound settings	Hardware Sync
UltraSuite	UXTD	Child	None	Research lab	58	Metal headset	Consistent	Correct
	UXSSD	Child	SSD	Research lab	8	Metal headset	Consistent	Correct
	UPX	Child	SSD	Research lab	20	Metal headset	Consistent	Correct
TaL	TaL1	Adult	None	Hemi-anechoic chamber	1	UltraFit headset	Consistent	Correct
	Tal80	Adult	None	Hemi-anechoic chamber	81	UltraFit headset	Consistent	Correct
Cleft		Child	Cleft	Hospital	29	AAA headset or handheld	Varied	Poor

Table 1: Data overview.

The children were aged 7-11 years at the time of data collection. Each child had either cleft palate only (CP), unilateral cleft lip and palate affecting one side of the lip and palate (UCLP), or bilateral cleft lip and palate (BCLP) affecting both sides. Some speakers had syndromes often associated with cleft lip and palate, including Stickler Syndrome, Treacher Collins Syndrome, and Pierre Robin Sequence. One child had an Adenoidectomy and a Tonsillectomy, and another one had scoliosis at the base of their skull. These medical conditions can lead to additional anatomical differences affecting the mandible, which make it challenging to acquire clear ultrasound images. This, combined with the often more severe nature of speech disorders associated with cleft lip and palate make the data more challenging to automatically process than previous datasets, such UltraSuite and TaL.

The data was recorded over a maximum of two sessions: Assessment and Therapy. Recordings took place in a hospital, and audio recordings contained the speech of both the children and therapists. The majority of utterances were recorded in the midsagittal view, but some were recorded in the coronal view. We annotated the direction of the probe manually and release the annotation with the dataset. See Figure 2 for sample ultrasound images taken from the Cleft data.

Ultrasound was recorded with a Micro system. The frame rate varied between 80-170 fps, and the field of view varied between 90-80°. The number of scan lines varied between 44-64, and the echo returns varied between 842-946. For the majority of speakers, the probe was stabilised with the AAA headset, but for two speakers (speaker 3 and 12), it was hand-held. Audio was recorded at 22.05 KHz sampling frequency.

We exported the data from AAA's proprietary format into the same format as UltraSuite and TaL. Four files are associated with each utterance. The **prompt** file is a *.txt* file containing the prompt the child was given and the date and time of the recording. The **waveform** is *.wav* file sampled at 22.05 KHz. **Ultrasound** data is stored as a matrix in a *.ult* file and is accompanied by a *.param* text file containing the metadata, such as frame rate, ultrasound frame size, and original hardware synchronisation offset. We complement this data with exported annotation from speech and language therapists in Praat's TextGrid format.

Utterance type	Type ID	Sagittal	Coronal	Total
Words	A	303	0	303
Non-words	B	503	73	576
Sentence	C	344	126	470
Non-Speech	E	49	43	92
All		1199	242	1441

Table 2: The number of utterances of each type in the Cleft dataset.

We categorised utterances into four types according to the prompts. **Words** contain a group of words designed to sample consonants in different vowel contexts within real words (e.g., "a core, a sip, a cop, a tool"). **Non-words** are designed to elicit specific phones but are not real English words (e.g., "acha" for /ʧ/). Many of these utterances contain multiple repetitions of the the same word (e.g., "acha acha acha acha"). **Sentences** are designed to examine specific phones in different contexts at the sentence level (e.g., "Tiny Tim is putting a hat on" for the phone /t/). And finally, **non-speech** utterances are swallowing motions recorded to trace the hard palate. We append the type ID to the utterance name (e.g., "001E.wav"). Table 2 summarises the data.

3.4. Cleft data challenges

A number of factors make the Cleft dataset more challenging to automatically process than TaL and UltraSuite, leading us to class it as a *new domain*. Firstly, the data was recorded with a clinical subgroup with severe speech disorders making audio more challenging to understand than the disordered subset of UltraSuite (UPX and UXSSD). Cleft patients also exhibit abnormal lingual articulatory patterns which are captured in ultrasound (Zharkova, 2013), and which will be different to patterns exhibited in UltraSuite and TaL. Furthermore, the anatomical differences arising from cleft lip and palate, as well as the additional syndromes that affect some of the children, can give rise to differences in the ultrasound data and in

Participant ID	Profession	Native English	Dialect
1	SLT	Yes	Scottish
2	Speech Scientist	Yes	British other
3	Speech Scientist	Yes	Scottish
4	Speech Scientist	No	Fluent, non-native
5	SLT	Yes	Scottish
6	SLT	Yes	Non-British
7	Speech Scientist	No	Fluent, non-native
8	SLT	Yes	British other
9	Speech Scientist	No	Fluent, non-native
10	Speech Scientist	No	Fluent, non-native

Table 3: Details of the participants.

some cases make it more challenging to acquire clear data in the first place.

Secondly, the data was recorded in a hospital environment with a lot of background noise. In contrast, UltraSuite was recorded in a quieter research laboratory, while TaL was recorded in a silent hemi-anechoic chamber.

Finally, the ultrasound in the Cleft data was recorded at varied settings including different frame rates, fields of view, scan lines, and echo returns, compared to the UltraSuite and TaL datasets which were consistent across speakers. Furthermore, the ultrasound probe was not always stabilised with a headset, leading to further inconsistency in the data. For these reasons we class the Cleft dataset as a new domain.

Because the Cleft data was poorly synchronised at recording time, we restrict its use to Section 6 where we automatically synchronise it using our system. In the next section, we examine the tolerance of expert users to synchronisation errors.

4. Identifying the detection threshold

This section aims to identify the threshold at which a synchronisation error becomes detectable to experienced ultrasound users. Identifying this threshold allows us to refine our approach for evaluating our system in Section 5. Because the movement of the articulators (the tongue and lips) are correlated, we turn to a study carried out with human participants which reports 6 different thresholds for lip synchronisation (ITU-R, 1998). We test whether the lip thresholds also apply to ultrasound in perceptual experiment, which we describe below[1].

4.1. Experiment

The purpose of this experiment was to discover how sensitive experienced ultrasound users are to different synchronisation errors. To this end, we recruited a number experienced ultrasound users, and asked them to assess the quality of audiovisual synchronisation in a series of recordings. During the experiment, we gave each participant pairs of videos containing ultrasound tongue imaging and the corresponding audio, and asked them to choose the videos which they perceive to be better synchronised. Each pair of videos were identical apart from the synchronisation offset. For one of the videos, we use the **correct hardware synchronisation offset**. For the other video, we **added an error to the correct offset**. The order of the videos was randomised, and the correct choice was unknown to the participants. We asked the following question: "In which of the two videos are the audio and tongue motion better synchronised, A or B?", and gave 3 choices: "Video A", "Video B", and "No perceived difference". We refer to the last as option C. We encouraged participants to make a choice between videos A and B, and to reserve option C for the most challenging cases. In this setting, the smaller the error the more challenging the task, and therefore, we expect the accuracy of choice to approach 50% when the error is imperceptible, and 100% when the error is perceptible.

The experiment was computer-based, and the videos were displayed on the participants' screens. The overall experiment lasted 30-40 minutes, and participants were allowed to complete it over multiple sessions. All utterances were in the midsagittal orientation with the tip of the tongue to the right. Ultrasound tongue imaging users typically playback ultrasound at three possible speeds: 1.0×, 0.5×, and 0.25×. We replicated this setting by giving our participants the option to play the videos at these three speeds. Participants were required to play each video at least once and up to 6 times at any speed, and could only move to the next video after they had submitted a judgement. To qualify, each participant was required to be a fluent English speaker, and either a speech and language therapist or a speech scientist with experience working with ultrasound tongue imaging. We recruited 10 participants whose details we outline in Table 3.

4.2. Data preparation

To test synchronisation errors, we required correctly-synchronised data. We therefore used the typically developing subset of UltraSuite, UXTD, which was correctly synchronised at recording time using hardware synchronisation. We chose this subset of UltraSuite to avoid distracting our participants with speech sounds disorders. The TaL corpus was not used for this experiment, as it was still in the process of being collected.

[1]This study was certified according to the Informatics Research Ethics Process (ref no 2019/43362).

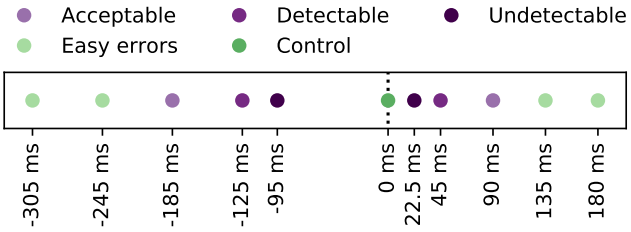

Figure 3: The set of synchronisation errors tested in our experiment. Audio lags when the error is negative and leads when positive. We tested the standard lip synchronisation error thresholds from ITU-R (1998) (acceptable, detectable, undetectable). The asymmetry indicates that errors are more challenging to detect when audio lags (negative) and easier to detect when audio leads (positive). We further tested four easy errors and add a control of zero error.

To get a rough idea of the audio quality, we listened to a small sample of audio recordings from each of the 58 speakers, then retained 42 speakers with the fewest interruptions from the therapists, fewest hesitations, and fewest deviations from the prompts. We sorted the speakers by the number of utterances, then by the standard deviation of the duration of utterances and chose the top 13 speakers (6 female, 7 male). These were speakers 1, 5, 6, 7, 9, 13, 17, 19, 20, 22, 23, 27, and 30. We selected a variety of prompts excluding coughs, and swallows and limited our selection to utterances shorter than 7.5 seconds. In total, we ended up with 520 unique recordings.

Next, we selected the set of errors to test, using the thresholds for lip synchronisation from ITU-R (1998). **Lip synchronisation errors** are classed as:

1. **Acceptable**: between -185ms and 90ms
2. **Detectable**: at -125ms and at 45ms
3. **Undetectable**: between -95ms and 22.5ms

Note the asymmetry in these thresholds: the magnitude of each positive error is smaller than its negative counterpart, indicating that errors are easier to detect if the audio leads, and more challenging to detect if the audio lags.

In earlier iterations of the experiment, we discovered that these thresholds were very challenging for our participants. Therefore, to make the experiment more engaging, and to give the participants less challenging cases to calibrate their answers to, we added four larger errors, two positive, and two negative. We selected these errors by computing the difference between the detectable and acceptable lip error thresholds, and used this difference to create two new evenly-spaced errors. We did this independently for positive and negative errors. Finally, we added a control, an error of zero. In this case, there was no difference between the pair of videos. The reason for adding this case is to test whether there is a bias towards choice A or choice B. For example, always preferring the video at the top of the screen would be a kind of bias. The final set of errors we tested is: [-305, -245, -185, -125, -95, 0, 22.5, 45, 90, 135, 180] ms (illustrated in Figure 3).

To create samples for our experiment, we randomly assigned the errors to the utterances. We drew 500 utterances from our pool of 520 and distributed them among the errors. We assigned each error 50 unique utterances, with the exception of the two most challenging errors, -95 and 22.5 (undetectable lip error) which we assigned only 25 each to avoid frustrating participants. Each participant evaluated 60 samples, 40 unique to them, and 20 shared with another participant to allow us to calculate pairwise agreement. In total, 500 unique samples were evaluated: 400 by a single participant, and 100 twice by a pair of participants, bringing the number of samples to 600. Each participant evaluated the same number of samples for each error. We report the results below.

Subset	Samples	Accuracy	CI
All	600	74.0%	(70.5, 77.5)
Excluding control	540	78.3%	(74.9, 81.8)
A is correct	262	77.9%	(72.8, 82.9)
B is correct	278	78.8%	(74.0, 83.6)
Control (C is correct)	60	35.0%	(22.9, 47.1)

Table 4: Overall accuracy of participant responses. CI are 95% binomial confidence intervals. Control pairs have a zero error for both videos. The percentages of A and B choices are similar indicating no bias towards A or B.

Error sign	Samples	Accuracy	CI
Negative	270	78.9%	(74.0, 83.8)
Zero (control)	60	35.0%	(22.9, 47.1)
Positive	270	77.8%	(72.8, 82.7)

Table 5: Overall accuracy of participant responses by sign. CI are 95% binomial confidence intervals. The accuracy is symmetrical despite the errors being asymmetrical indicating that the lip synchronisation asymmetry also holds for ultrasound tongue imaging. Negative: audio lags. Positive: audio leads. Zero: no error (control).

4.3. Results

The first results are shown in Table 4. The overall accuracy of participant choice was 74.0%. For control questions, where both videos had no synchronisation error, participants selected C only 35.0% of the time. As for non-control questions, participants chose the correct answer 78.3% of the time. The percentages of A and B choices were balanced (77.9% and 78.8%) indicating no bias in choice towards A or B. Table 5 displays the accuracy by sign. The table shows that accuracy is sym-

Category	P	N	Accuracy	CI
SLT	4	216	85.6%	(81.0, 90.3)
Speech Scientist	6	324	73.5%	(68.6, 78.3)
Fluent, non-native	4	216	70.8%	(64.8, 76.9)
Scottish	3	162	79.6%	(73.4, 85.8)
British other	2	108	88.9%	(83.0, 94.8)
Non-British	1	54	83.3%	(73.4, 93.3)

Table 6: Accuracy of participant responses excluding control, broken down by the participants professions (top) and their dialects (bottom). P is the number of participants, while N is the number of samples. CI are 95% binomial confidence intervals.

metrical despite the errors being asymmetrical, indicating that the asymmetry that holds for lip synchronisation also holds for ultrasound tongue imaging.

Figure 4 breaks the accuracy down by participant and by error. As expected, the smaller the error, the more challenging the task. The confidence intervals for the undetectable lip error thresholds both cross 50%. The confidence intervals for 45ms reaches 50%, indicating that even the detectable lip error thresholds are too challenging for ultrasound tongue imaging. We start to see more reliable accuracy at the acceptable lip error thresholds. Finally, Figure 5 shows the percentage of C choices, or "no perceived difference".

Next, we calculated pairwise agreement. Each pair of participants (1 & 2, 3 & 4, ...etc.) received a common subset of 20 samples. The synchronisation errors had an equal number of common samples, 20 each, with the exception of undetectable lip errors, -95 and 22.5 which had 10 samples each. We calculated the following **scores for pairwise agreement**:

1. **Agreement of choice**: did the participants make the same choice (A, B, or C)?
2. **Agreement of outcome**: did their choice have the same outcome (both correct or both incorrect)?
3. **Agreement with truth**: did the choice match the truth (both correct)?

Figure 6 shows the results by participant pair and by synchronisation error. All pairs of participants agreed on at least 50% of samples. As expected, the smaller the error, the lower the agreement, with the exception of the undetectable error at -95ms and 22.5ms, where agreement is lower than expected at -95ms and higher than expected at 22.5ms, possibly due to the smaller sample size. Another contributor could be the randomisation procedure: because utterances were randomly assigned errors, it is possible that certain errors had easier / more challenging utterances by chance.

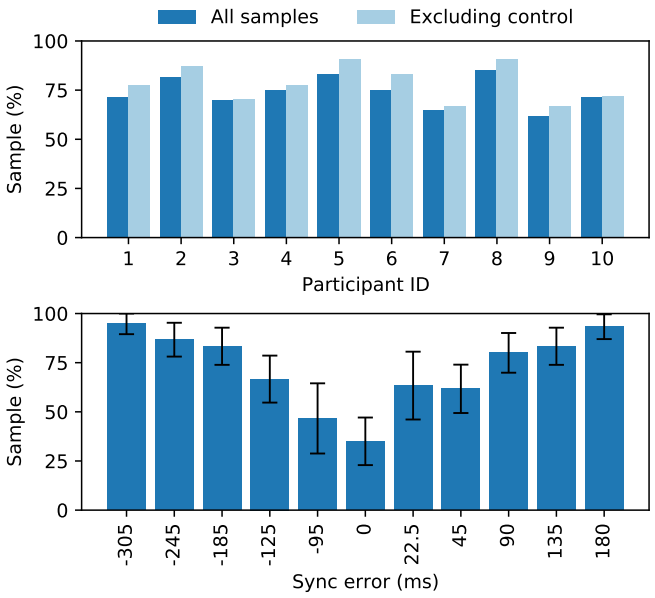

Figure 4: Accuracy of choice shown by participant (top) and by synchronisation error (bottom). The smaller the error, the more challenging the task. The confidence intervals for the undetectable lip errors cross 50% indicating that they are also undetectable for ultrasound. The confidence intervals for 45ms also reaches 50%, indicating that this threshold for detecting lip error is not applicable to ultrasound. The accuracy at the threshold for acceptable lip error is more reliable.

English Speaker	Profession	P	N	Accuracy	CI
Native	SLT	4	216	85.6%	(81.0, 90.3)
Native	Speech Scientist	2	108	78.7%	(71.0, 86.4)
Non-native	Speech Scientist	4	216	70.8%	(64.8, 76.9)

Table 7: Accuracy of participant responses excluding control, split by the combination of native language and profession. CI are 95% binomial confidence intervals. Native English-speaking SLTs perform the task better than non-native English-speaking speech scientists. The CI of the middle group (native English speaking speech scientists) overlaps with the two other groups.

To understand why the overall accuracy varied by participant, we broke the results down by their professions and dialects in Table 6. Four participants were speech and language therapists (SLTs) while six were speech scientists. As for their dialects, 4 were fluent non-native English speakers and 6 were native English speakers: 3 Scottish, 2 non-Scottish British, and 1 non-British. Table 6 shows that SLTs achieved higher accuracy than speech scientists, however, Table 7 shows that the profession of participants co-varied with their native language, and that not all combinations are represented in out data. For example, all non-native English speakers were speech scientists and none were SLTs. While such characteristics may have an effect on a user's sensitivity to synchronisation offsets, from our data it is not possible to isolate the

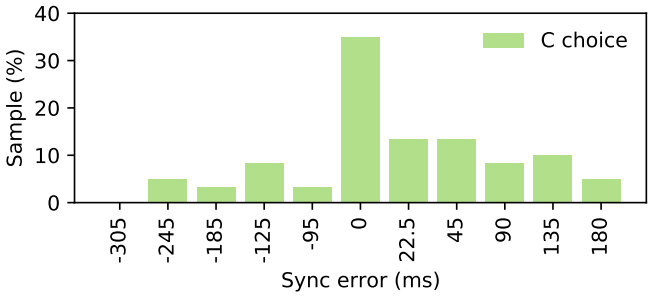

Figure 5: The distribution of C choices ("no perceived difference") per synchronisation error. The majority of C choices are concentrated at zero error (where there is in fact no difference), and the distribution tapers as the errors become larger.

individual effects of profession and native language.

Finally, we conducted a linear analysis, predicting the outcome of choice (correct/incorrect) from the synchronisation error while controlling for the participant and utterance content. We represented errors and participants as one-hot encoding vectors, and introduced content features at the phone level to test whether synchronisation errors are easier to detect in the presence of certain phones. To map each word to its pronunciation, we used the UXTD pronunciation dictionary supplied with the data. The pronunciation dictionary was compiled for a Scottish accent (to match the accent in the data) using the Combilex lexicon (Richmond et al., 2010, 2009). We found 42 unique phones in the test utterances. For each utterance, we created a feature vector of size 42, and counted the number of occurrences for each phone. For words with multiple pronunciations, we added fractional counts for each phone as $Count = \frac{1}{P}$, where P is the number of pronunciations.

We then fit a logistic regression model predicting the binary outcome (correct / incorrect) from 63 features: 11 errors, 10 participants, and 42 phones. We used LBFGS with $L2$ regularisation. Upon convergence, the model achieved a log loss of 0.456. To calculate the proportion of model variation that is explained by the features, we used McFadden's pseudo-R^2. The score falls between 0 and 1, however, in practice, scores ranging from 0.2 to 0.4 are considered excellent (Hensher & Stopher, 1979) and indicate that a large proportion of the model is explained by the features. Our model's pseudo-R^2 score is 0.249.

The model coefficients are shown in Figure 7. The direction of the coefficients (positive / negative) is the direction of correlation with correctness of participant choice. We find that the tolerance for synchronisation error varies by participant. Synchronisation between -125 and 45 ms negatively correlate with a correct choice. For lip synchronisation, these are the thresholds for detection. However,

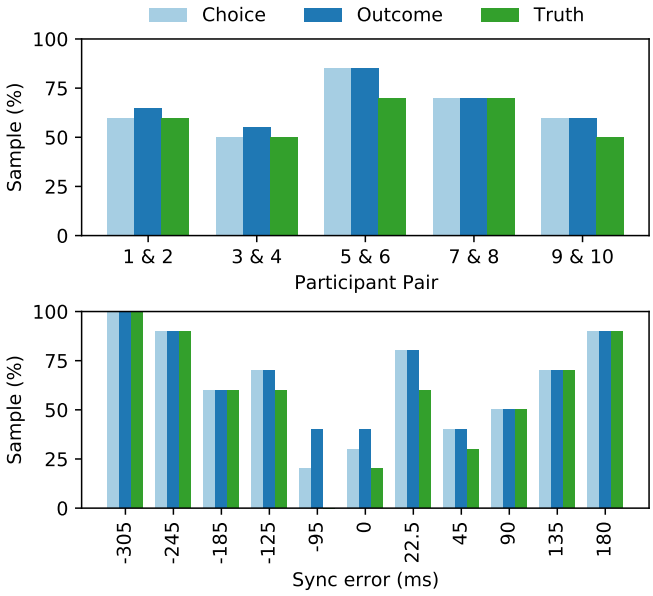

Figure 6: Pairwise agreement shown by participant pair (top) and by synchronisation error (bottom). Each participant pair shared 20 samples and agreed on at least half of them. Each error had 20 samples except errors -95ms and 22.5ms which had only 10 each. The smaller sample size might explain why agreement is lower than expected for -95ms and higher than expected for 22.5ms. Otherwise, agreement positively correlates with error magnitude.

these results indicate that for ultrasound, undetectability extends to this range. Errors <=-185 ms and errors >=90 ms positively correlate with a correct choice. We therefore define the following thresholds for **ultrasound synchronisation errors**:

1. **Detectable**: at -185ms and at 90ms
2. **Undetectable**: between -125ms and 45ms

and use them in Sections 5 to evaluate our system.

Because we represented phone as fractional counts, and represented participants and errors as one-hot vectors, the magnitudes of coefficients are not directly comparable. However, the direction of the coefficients is the direction of correlation. The results for utterance content meet our expectations. Phones that involve little tongue movement, such as those produced using the lips (for example /b/) or the glottis (for example /h/), negatively correlate with a correct choice. In contrast, phones involving more tongue activity (alveolars, post alveolars, palatals, and velars) positively correlate with a correct answer. This result is intuitive and meets our expectations.

4.4. Discussion and summary

In this section, we applied the standard lip error thresholds to ultrasound and tested them in a perceptual experi-

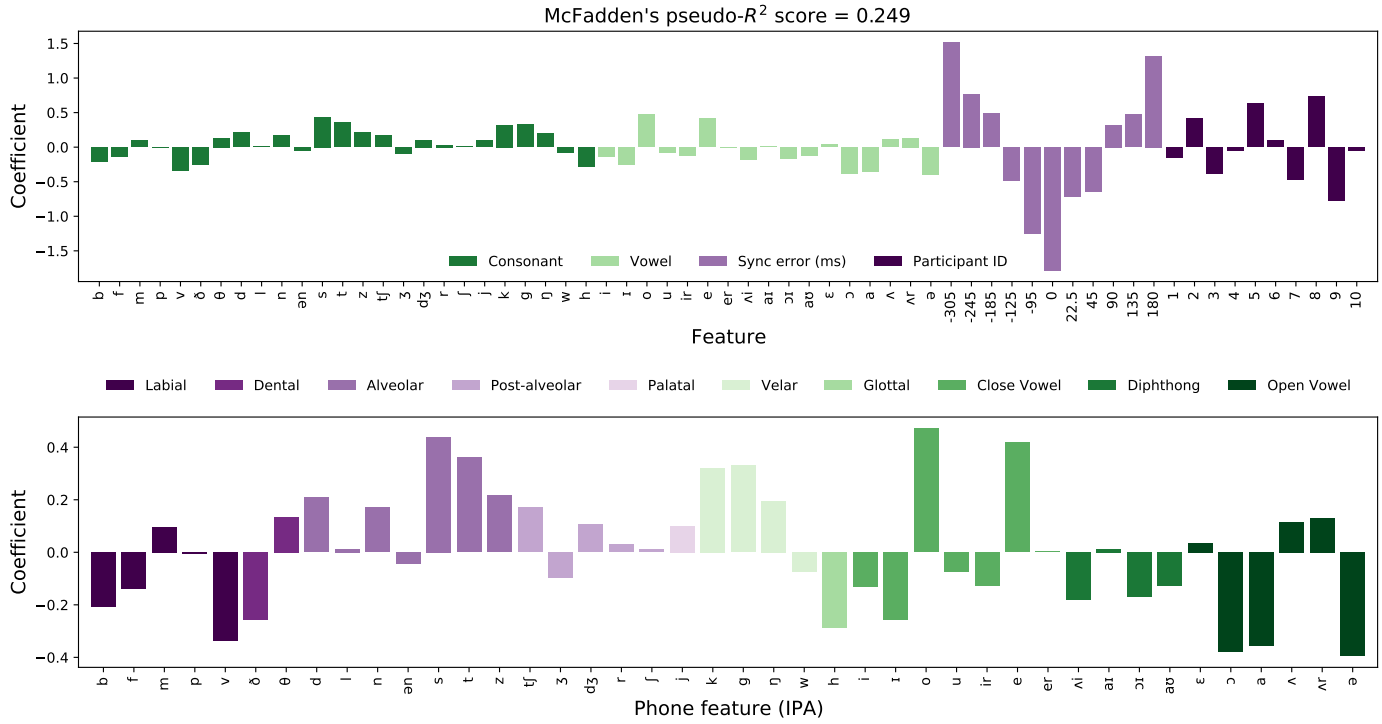

Figure 7: The coefficients of a logistic regression model predicting choice outcome (correct / incorrect) from the features shown. McFadden's pseudo-R^2 score shows the proportion of model variation explained by the features. **Top:** the complete set of features. Synchronisation errors falling between −125ms and 45ms negatively correlate with a correct choice, while errors <= −185ms and errors >=90ms positively correlate with a correct choice. Sensitivity to synchronisation errors varies by participant. **Bottom:** the phone features colour-coded by the place of articulation. Alveolars, post-alveolars, palatals, and velars positively correlate with a correct choice, while labials, dentals, and glottals negatively correlate with a correct choice. Most vowels (close vowels, open vowels, and diphthongs) negatively correlate with a correct choice, with a few exceptions such as 'o' and 'e'.

ment with expert ultrasound users. We concluded that detecting synchronisation errors in ultrasound tongue imaging is more challenging than in lip videos. This is perhaps not surprising given that most humans are exposed to audiovisual perception of lip movement from birth, therefore accumulating thousands of hours of experience seeing synchronised lips and audio. The same does not hold for ultrasound; even the most experienced ultrasound users only have tens or hundreds of hours of experience working with synchronised ultrasound and audio. Moreover, ultrasound images, unlike videos of lips, are not a facsimile, or indeed even a video, instead they are a representation of tongue-movements based on echos of high-frequency sound waves and as such are susceptible to artefacts. It is therefore reasonable for the synchronisation error detection threshold to be larger than for lip videos.

We further concluded that the sensitivity to synchronisation errors varies by participant, after taking into account the linguistic content of utterances and the offsets as co-variates in the linear model.

Finally, we concluded that sounds involving high tongue activity positively correlate with synchronisation error detection, while sounds involving low tongue activity negatively correlate with synchronisation error detection.

5. Automatic synchronisation system

This section details our approach for automatically synchronising ultrasound and audio. We build directly on our model from Eshky et al. (2019) reiterating its description below. We then describe a new experiment, introduce two evaluation scores based on the results from Section 4, and present our results on in-domain data.

5.1. Model

We use the UltraSync architecture from Eshky et al. (2019) which previously extended the work of Chung & Zisserman (2016) on lip synchronisation, modifying for ultrasound tongue imaging. The system accepts as input an ultrasound signal and an audio signal, and requires the range of possible offsets to be specified. From this range, the system selects the offset that minimises the distance between the two signals.

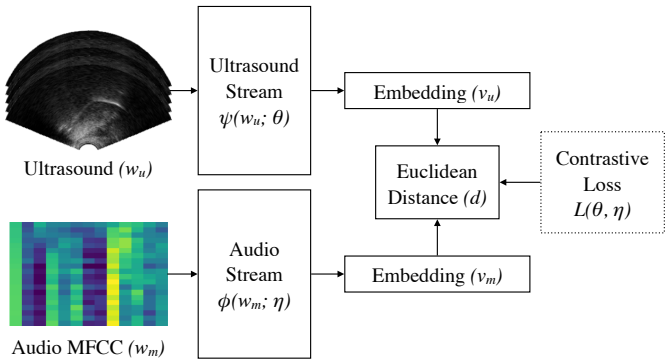

Figure 8: The UltraSync model accepts as input a window of ultrasound and a window of audio, represented as MFCC features. Each stream is a convolutional neural network mapping the inputs to low dimensional embeddings. The model then outputs the Euclidean distance between the embeddings, and a contrastive loss function minimises the distance for true pairs and maximises it for false pairs.

At the heart of the system is a neural network with two streams, illustrated in Figure 8. The first stream accepts a short window of ultrasound, and the second accepts a short window of audio. The inputs are of different sizes and are high-dimensional. The network maps the pair of inputs to a pair of low-dimensional embeddings of equal lengths, such that the Euclidean distance between them is small when they correlate and large otherwise.

The learning objective is a contrastive loss function (Chopra et al., 2005; Hadsell et al., 2006), which minimises the Euclidean distance between embeddings from "true" input pairs, and maximises it for "false" input pairs. True and false pairs are automatically generated from the training data through a process known as self-supervision.

Formally, the network maps a window of ultrasound w_u, and a window of audio w_m (represented as MFCC features), to two low dimensional embeddings v_u and v_m:

$$\begin{aligned} \psi(w_u; \theta) &\rightarrow v_u \\ \phi(w_m; \eta) &\rightarrow v_m \end{aligned} \tag{1}$$

Where ψ and ϕ are non-linear transformations with parameters θ and η. The network then calculates the Euclidean distance d between the embeddings:

$$d = \|v_u - v_m\|_2 \tag{2}$$

The learning objective is a contrastive loss L, which minimises the distance d for true pairs (labelled $y = 1$), and maximises it for false pairs (labelled $y = 0$), for a number of training samples N:

$$L(\theta, \eta) = \frac{1}{N} \sum_{n=1}^{N} y_n d_n^2 + (1 - y_n)\{max(1 - d_n, 0)\}^2 \tag{3}$$

Algorithm 1: Synchronisation algorithm

```
Input: ultrasound, audio, and candidate offsets
for each candidate do
    Apply candidate to utterance
    Create windows of ultrasound and audio
    for each window do
        Calculate the distance between ultrasound
          and audio using UltraSync
    end
    Calculate the mean distance
end
Select the offset with the smallest mean distance
```

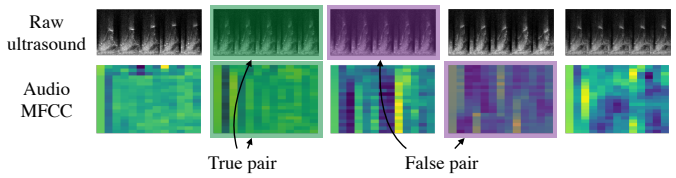

Figure 9: We create training samples automatically using a self-supervision strategy. For each utterance, we create short windows of ultrasound and audio. True samples are corresponding pairs, and false samples are randomised pairings. Ultrasound frames are shown as raw reflection data.

Once trained, the model can be used to calculate the Euclidean distance between a pair of ultrasound and audio windows.

To find the synchronisation offset, we first need to specify the range of possible shifts (e.g., ±1000 ms). Within this range, we use our model to identify the offset that minimises the mean Euclidean distance across shorter windows of the two signals. In practice, we discretise the range of possible shifts, rendering a discrete set of candidate offsets. Then, using Algorithm 1, we calculate the mean euclidean distance for each of these candidates, and select the one with the smallest mean distance as our prediction.

To train our model, all we require is a training set with correctly synchronised utterances, and from this dataset we automatically generate true and false pairs. From each utterance in the set, we generate multiple true pairs by creating short windows of ultrasound and corresponding audio and labelling them as true. To create false pairs, we simply randomise the pairings within each utterance, and label them as false. Figure 9 illustrates the process of creating true and false samples.

5.2. Experiment

For this experiment, we used the UltraSuite and the TaL data. The datasets were recorded with speakers with dif-

ferent characteristics, in different environments, and using different equipment. We utilised such data to enable our model to accommodate different speakers groups, recording conditions, and ultrasound probe types. We split UltraSuite and TaL into training, validation, and testing subsets. We used the same data splits for UltraSuite as Eshky et al. (2019), and the same data splits for TaL as Ribeiro et al. (2021b) for comparability. We reiterate the data splits below.

From UXTD, we used speakers [7, 8, 12, 13, 26] for validation, [30, 38, 43, 45, 47, 52, 53, 55] for testing, and all remaining 45 speakers for training. From UXSSD, we used speaker 1 and session 'Mid' from speakers [2, 3, 4] for validation, speaker 7 and session 'Mid' from speakers [5, 6, 8] for testing, and all remaining speakers and sessions for training. From UPX, we used speaker 1 and session 'BL3' from speakers [2-10] validation, speaker 15 and session 'BL3' from speakers [11-14] and [16-20] testing, and all remaining speakers and sessions for training. We used utterances containing words, non-words, sentences, isolated articulations, and conversations, and excluded utterance containing only coughs and swallowing motions.

From TaL1, we used days [2, 3, 4] for training, day 5 for validation, and days [1, 6] for testing. From TaL80, we used speakers [1-49] for training, speakers [50-65] for validation, and speakers [66-81] for testing. We used read and spontaneous utterances and excluded swallow, silent, and whispered utterances.

A pre-processing step re-sampled audio at 22.05 KHz (using *scipy interpolate*), re-sampled ultrasound at 24fps (using *samplerate resample*), and resized ultrasound frames to 63×138 pixels (using *skimage transform*).

We defined the sample window size as ≃200ms long, calculated as $t = l/r$, where t is the time window, l is the number of ultrasound frames per window (5 in our case), and r is the ultrasound frame rate of the utterance (24 fps). For each utterance, we split the ultrasound into non-overlapping windows of 5 frames each. To create corresponding audio windows, we extracted MFCC features from the audio signal, with 13 cepstral coefficients, using a window length of ≃20ms, calculated as $t/(l \times 2)$, and a step size of ≃10ms, calculated as $t/(l \times 4)$. We chose MFCCs as they are one of the most frequently used representations in the speech processing literature, and have been shown to work for lip video synchronisation (Chung & Zisserman, 2016). We created true and false training samples using the process outlined in Figure 9, and generated as many false pairs as true ones for a balanced set.

The hyper-parameters of our network are shown in Table 8. We pooled all training data and trained a single model using the Adam optimiser (Kingma & Ba, 2015), with a learning rate of 0.001, a batch size of 64 samples, and for 20 epochs. We implemented learning rate scheduling, which reduced the learning rate by a factor of 0.1 when the validation loss plateaued for 2 epochs.

Stream	Conv	Conv	Conv	FC	FC
Ultrasound $5 \times 63 \times 138$	$23 \times 5 \times 5$ $\times 2$ pool	$64 \times 5 \times 5$ $\times 2$ pool	$128 \times 5 \times 5$ $\times 2$ pool	64	64
Audio $1 \times 20 \times 30$	$23 \times 3 \times 3$	$64 \times 3 \times 3$ $\times 2$ pool	$128 \times 3 \times 3$ $\times 2$ pool	64	64

Table 8: Each stream had 3 convolutional (Conv) layers followed by 2 fully-connected (FC) layers. FC layers had 64 units each. For Conv layers, we specify the number of filters and their receptive field size as "num × size × size" followed by the max-pooling down-sampling factor. Each layer was followed by batch-normalisation then ReLU activation. Max-pooling was applied after the activation function.

Upon convergence, the model achieved 0.19 training loss, 0.19 validation loss, and 0.20 test loss, and by placing a threshold of 0.5 on predicted distances, the model achieved 71.7% binary classification accuracy on training samples, 71.3% on validation samples, and 69.3% on test samples.

5.3. Evaluation and results

Next, we followed Algorithm 1 to predict the offsets for the test utterances, using the same 24 candidates for UltraSuite as Eshky et al. (2019), and using the same 25 for TaL as Ribeiro et al. (2021b).

To evaluate the predictions, we computed the discrepancy between the model prediction and the true offset as:

$$disc = prediction - truth \tag{4}$$

Because hardware synchronisation was correct for UltraSuite and TaL, we treat it as *truth*. We consider the prediction to be correct if it falls between *lower* and *upper* thresholds:

$$lower < disc < upper \tag{5}$$

Based on the new threshold defined in Section 4, we define two **accuracy scoring boundaries**:

1. **Hard**: *lower* = −125ms and *upper* =45ms
2. **Soft**: *lower* = −185ms and *upper* =90ms

The hard scoring boundary is the same one used in previous work on lip synchronisation (Chung & Zisserman, 2016) and ultrasound synchronisation (Eshky et al., 2019; Ribeiro et al., 2021b). However, in Section 4, we found

Subset	N	Hard	Soft	Discrepancy
UltraSuite: child data				
UXTD	455	64.6%	74.5%	123 ± 392 ms
UXSSD	396	88.9%	95.7%	12 ± 146 ms
UPX	651	93.7%	98.3%	0 ± 90 ms
	1502	83.6%	90.4%	41 ± 242 ms
TaL: adult data				
TaL1	452	98.7%	99.8%	0 ± 26 ms
TaL80	3129	95.7%	98.2%	-8 ± 54 ms
	3581	96.1%	98.4%	-7 ± 51 ms
All	5083	92.4%	96.0%	7 ± 140

Table 9: Results by dataset. We show the accuracy using hard and soft scoring boundaries, and the mean and standard deviation of the discrepancy in milliseconds. Performance on adult data (TaL) is better than on child data (UltraSuite).

Utterance Type	N	Hard	Soft	Discrepancy
UltraSuite: child data				
Words	914	92.0%	97.7%	3 ± 107 ms
Non-words	58	86.2%	98.3%	14 ± 165 ms
Sentence	186	94.6%	97.3%	11 ± 105 ms
Articulatory	340	54.4%	65.6%	164 ± 445 ms
Conversation	4	100%	100%	-20 ± 19 ms
TaL: adult data				
Read	2979	95.8%	98.2%	-8 ± 54 ms
Read shared	432	97.7%	99.3%	-2 ± 36 ms
Spontaneous	18	94.4%	100%	-9 ± 32 ms
Calibration	152	98.7%	100%	-3 ± 18 ms

Table 10: Results by utterance type. We show the accuracy using hard and soft scoring boundaries, and the mean and standard deviation of the discrepancy in milliseconds. Articulatory utterances contain isolated phones and are the most challenging. In contrast, performance is high on utterances containing natural variation in speech, such as words, sentences, conversations, read text, and spontaneous speech.

these thresholds to be too strict for ultrasound, and so we also present results using the soft scoring boundary.

Table 9 shows the results by dataset. The model correctly synchronises 92.4% of utterances according to the hard scoring boundary and 96.0% of utterances according to the soft scoring boundary. The overall discrepancy is 7 ± 140 ms. Performance on TaL is better than on UltraSuite. On child data (UltraSuite), the model achieves an overall hard accuracy of 83.6%, a marginal improvement of 0.7% over Eshky et al. (2019), and achieves a soft accuracy of 90.4%. On adult data (TaL), the model achieves an overall hard accuracy of 96.1%, a marginal reduction of 1.6% over Ribeiro et al. (2021b), and achieves a soft accuracy of 98.4%. While these differences are small, they make intuitive sense. UltraSuite was recorded during speech therapy sessions in noisy environments, and the audio contains the speech of both therapists and patients. TaL on the other hand, was recorded in a hemianechoic chamber to eliminate background noise, and the audio and ultrasound always corresponded to the same speaker, resulting in much better overall quality. Therefore, it is unsurprising that training on TaL improves the performance on UltraSuite, while training on UltraSuite reduces the performance on TaL.

Table 10 shows the results by utterance type. Performance according to both scoring boundaries is highest on utterances containing natural variation in speech, such as words, sentences, read text, and spontaneous speech. This result is consistent with the results from Eshky et al. (2019). Articulatory utterances, on the other hand, contain isolated phones (e.g. sh), and therefore lack natural variation in speech, which makes them more challenging to automatically synchronise. Nonetheless, the model correctly synchronises 54.4% of these utterances according to the hard scoring boundary, and 65.6% of the utterances according to the soft scoring boundary.

Non-word stimuli are designed to elicit phones in different contexts from patients, but are not real English words (e.g. “p apa epe opo”). To some extent, these utterances also lack natural variation in speech. According to the hard scoring boundary, 86.2% of these utterances are correctly synchronised, which is lower than the accuracy achieved for words and sentences. However, using the slightly more flexible soft scoring boundary, 98.3% of these utterances are considered correctly-synchronised, which slightly exceeds performance on words and sentences. At a first glace, this result seems surprising, but considering that many of these utterances contain repetitions of the same non-word, it is possible that the model is able to identify periodic landmarks in the utterances, and synchronise them to an adequate level, if not as precisely as it synchronises words and non-words.

To summarise, in this section we presented our approach for automatically synchronising ultrasound and audio. We introduced two scoring boundaries based on the detection thresholds from Section 4, and showed how to use them to evaluate our model. Results are consistent with previous work, demonstrating that performance is highest on utterances exhibiting natural variation in speech. TaL is of better quality than the UltraSuite data, and it is therefore unsurprising that the model achieves higher performance on TaL than on UltraSuite. Training a single model on the pooled TaL and UltraSuite data slightly reduces the performance on TaL and slightly in-

Participant ID	Profession	Native English	Dialect
1	SLT	Yes	Scottish
2	SLT	Yes	Non-British
3	Speech Scientist	Yes	British other
4	Speech Scientist	Yes	Scottish
5	SLT	Yes	Scottish
6	SLT	Yes	Non-British

Table 11: Details of the participants.

creases it on UltraSuite, compared to previous research. In the next section, we evaluate our model's performance on the out-of-domain Cleft data.

6. Synchronising the Cleft data

In this section, we test the performance of our system on the out-of-domain Cleft data. As described in Section 3, hardware synchronisation for the Cleft data was perceived as inadequate by the speech and language therapists who recorded it. Because correct synchronisation is not available for this data, we are unable to automatically evaluate model performance as we did in the Section 5. Instead, we utilise the judgement of experienced ultrasound users in a second perceptual experiment, which we describe below[2].

6.1. Experiment

The experimental setup is similar to that in Section 4 with some differences which we outline below. We recruited a number of experienced ultrasound tongue imaging users, giving them pairs of videos containing ultrasound tongue imaging and the corresponding audio, and asking them to choose the videos which they perceived to be better synchronised. Each pair of videos were identical apart from the synchronisation offset. For one of the videos, we used the **original hardware synchronisation offset**. For the majority of utterances, this was perceived as inadequate by the speech and language therapists who collected the data. For the other video, we used the **offset predicted by our model**. The order of the videos was randomised and the source of the offset for each video was not shown to participants. This setting allowed us to measure the percentage of utterances for which the model improved synchronisation.

As in Section 4, the experiment was computer-based, and the videos were displayed on the participants' screens. The overall experiment lasted 30-40 minutes, and participants were allowed to complete it over multiple sessions. We gave the participants the option to play the videos at three speeds: 1.0×, 0.5×, and 0.25×, and required them to play each video at least once and up to 6 times at any speed. The participants could only move to the next pair of videos after submitting a judgement.

We asked the following question: "In which of the two videos are the audio and tongue motion better synchronised, A or B?". Unlike the experiment in Section 4, we gave the participants only 2 choices: "Video A", and "Video B", and asked them to chose at random if they perceived no difference, or if the synchronisation in both videos was equally poor. In this setting, preference would approach 50% if all choices were at random, and 100% if one method was always preferred. To qualify, each participant was required to be a fluent English speaker, and either a speech and language therapist or a speech scientist with experience working with ultrasound tongue imaging. We recruited 6 participants whose details we outline in Table 11.

6.2. Data preparation

The Cleft dataset contains 1441 samples of approximately 4.1 hours of audio in total. We evaluated only a subset of this data. We focused on evaluating spoken utterances (these are types A, B, and C in Table 2) and excluded "swallows" (type E) which have almost no audible content. We evaluated utterances recorded during assessment, and excluded therapy utterances as they tend to be much longer and tend deviate from the prompt. Because the model was only trained on midsagittal utterances with the tip of the tongue to the right, we excluded utterances recorded in the coronal orientation. The duration of recordings in the dataset range from 2.4 to 40 seconds, with a mean of 10.3 seconds and a standard deviation of 5.1 seconds. We placed a threshold of <=15 seconds on utterances to evaluate, thereby excluding the tail of longer utterances. We further excluded all utterances where the difference between the offset predicted by our model and the hardware offset fell within the undetectable range.

As we did in the first experiment, we listened to a small sample of recordings from each speaker to assess the audio quality. Out of the 29 speakers, we excluded 8 speakers who repeatedly deviated from the prompts and had the most interventions from therapists, because these kind utterances would distract our evaluators from the main task. We used the following speakers (9 female and 12 male):

[2]This experiment was certified according to the Informatics Research Ethics Process (ref no 2019/43362).

Utterance type	Type ID	N	Preference	CI
Words	A	100	78.0%	(69.9, 86.1)
Non-words	B	100	86.0%	(79.2, 92.8)
Sentence	C	100	74.0%	(65.4, 82.6)

Table 12: Preference for our model, shown by utterance type. CI are 95% binomial confidence intervals.

3, 5, 7, 11, 14, 15, 16, 17, 18, 20, 21, 24, 25, 28, 30, 31, 32, 33, 34, 36, 39.

To apply our approach to the Cleft data, we needed to specify the range of offsets, as we did in Section 5. We observed that for the majority of Cleft utterances, audio is advanced with respect to ultrasound, and so we considered a wider range of negative offsets than positive ones. The range we considered was [-1.75, 0.75] seconds with a step size of 45ms. This rendered 56 candidate offsets for the model to consider. We ran the model and reviewed a sample of predictions. We observed that the utterances with extreme offsets (largest and smallest) were poorly synchronised compared to utterances in the middle range. At this point, we had the option to either fine tune the range of candidate offsets, or sample utterance from the middle range. We chose to do the latter, randomly sampling 100 utterances of each utterance type (A, B, and C) within offsets [-1.5, 0.5], or a total of 300 utterances.

To test the reliability of participant choices, we added a small number of control utterances for which correct synchronisation was known. We used the UPX subset of UltraSuite, selecting 10 utterance with similar prompts to the Cleft dataset to obscure the origin of the utterances. We then created pairs of videos, which were identical apart from the synchronisation offset. For one of the videos, we use the correct hardware synchronisation offset and for the other, we added a detectable error of -305 ms for half of the utterances and 180 ms to the other half. All participants evaluated this same subset of 10 control utterances. In total, each participant evaluated 60 utterances, 50 Cleft samples and 10 control samples.

6.3. Results

Figure 10 shows the aggregate result and the result by participant. Results show that participants are highly reliable, achieving an accuracy of 91.7% with a confidence interval of (84.7, 98.7) for control utterances. As for Cleft samples, participants preferred the model's prediction over hardware synchronisation 79.3% of the time, with a confidence interval of (74.8, 83.9). We conduct a two-sided binomial test, achieving a p-value of $1.81e^{-25} < 0.001$, which indicates that the difference between the synchronisation methods is significant. We therefore have

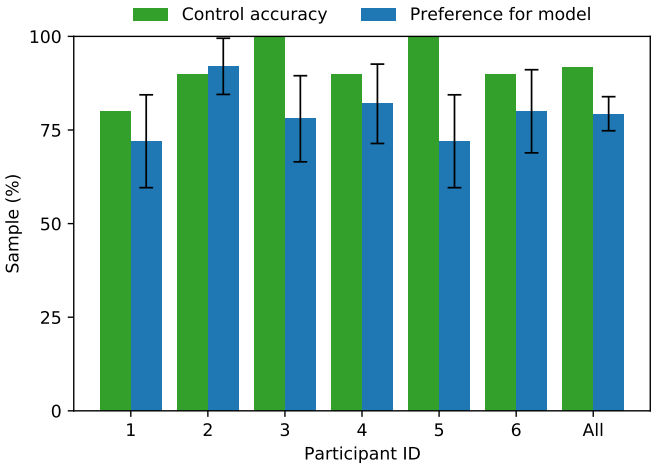

Figure 10: The aggregate result and the result by participant. Green bars show high choice accuracy for control samples, and therefore, high participant reliability. Blue bars show a strong preference for our model for Cleft samples. Confidence intervals are calculated for a binomial proportion.

Category	P	N	Preference	CI
SLT	4	200	79.0%	(73.4, 84.6)
Speech Scientist	2	100	80.0%	(72.2, 87.8)
Scottish	3	150	75.3%	(68.4, 82.2)
British other	2	100	79.0%	(71.0, 87.0)
Non-British	1	50	92.0%	(84.5, 99.5)

Table 13: Preference for our model, broken down by the participants' professions (top) and their dialects (bottom). P is the number of participants, while N is the number of samples. CI are 95% binomial confidence intervals.

sufficient evidence that participants prefer the output from our model over the original hardware synchronisation.

Table 12 shows the preference for our model broken down by utterance type. According to participant choice, our model performs best on utterances of type "non-words", followed by "words" then "sentences". As with the results in Section 5.3, this result may seem surprising at a first glance, as we expected performance to be higher on words and sentences because they exhibit slightly more natural variation in speech than non-words. However, the result is consistent with the Soft score calculated on in-domain data in Table 10. Because many of the "non-word" utterances contained repetitions of the same non-word (e.g., "aka aka aka.."), it is possible that poor synchronisation was more obvious, and easier to detect by our participants.

Finally, we break the results down by the professions and dialects of the participants in Table 13. Four of the participants were speech and language therapists (SLTs) and two were speech scientists. The results show no dif-

ference in model preference between the two groups. All participants were native English speakers: 3 Scottish, 2 non-Scottish British, and 1 non-British. The non-British speaker has a higher preference from our model, however due to the small sample size and the fact that the confidence intervals overlap with the non-Scottish British group, it is difficult to draw robust conclusions about the effects of dialect on model preference.

To summarise, in this section we applied our model to the Cleft data and evaluated its performance with experienced ultrasound tongue imaging users. The participants showed a strong preference for our model's output over hardware synchronisation, which demonstrates our model's ability to generalise to data from a new domain.

7. Conclusion

This paper addressed the problem of automatically synchronising ultrasound tongue imaging with speech audio. The two modalities are simultaneously-acquired; however, synchronisation information is not always correctly-captured at recording time, and is not always available for historical data.

In Section 4, we presented a novel investigation of the synchronisation errors tolerance by expert ultrasound users, and found that thresholds for error detection are greater for ultrasound tongue imaging than for lip videos. We also found that sensitivity to synchronisation errors varies by participant, and that phones involving little tongue movement negatively correlate with a correct choice, while phones involving more tongue activity positively correlate with a correct answer. Findings from this experiment allowed us to define thresholds for detecting synchronisation errors in ultrasound.

We then presented our approach for automatic synchronisation in Section 5, which utilises a self-supervised neural network to find the offset between ultrasound and audio in a given range. We defined two scoring boundaries for evaluating our model, a hard one and a soft one, based on the error thresholds we identified in our first perceptual experiment. We evaluated our approach in the first instance on in-domain data; a held-out subset of the data used to develop the model. Results are consistent with previous work, demonstrating that performance is highest on utterances exhibiting natural variation in speech. Our model achieved a higher performance on TaL than on UltraSuite, and training a single model on the pooled TaL and UltraSuite data slightly reduced accuracy on TaL, while improving it on UltraSuite, compared to previous research.

In Section 3, we introduced a novel resource, the Cleft dataset, which we collected with a new clinical subgroup, and for which hardware synchronisation proved unreliable. We applied our model to this data in Section 6 and evaluated it subjectively with expert users in a second perceptual experiment. We found that users preferred the output of our model 79.3% of the time, and that this result is statistically significant. These results demonstrate the strength of our model and its ability to generalise to new domains.

8. Discussion and future work

There are several avenues for future research. In Section 4, we investigated whether lip thresholds hold for ultrasound, and this served as a good starting point for identifying suitable thresholds to use for evaluating our system in Section 5. In the future, we can use the thresholds that we have identified as a guide to a new experiment which tests more fine-grained offsets to find more precise error detection thresholds.

Furthermore, in Section 4, we explored the notion of synchronisation error detection, but did not explicitly address "acceptable" error, simply because it depends on the end task. As discussed in Section 2.2, speech and language therapists use ultrasound differently to phoneticians, and so different tasks may require different levels of synchronisation precision. One future direction is to examine the effect of synchronisation error on the performance of experts in a downstream task, such as correctly identifying covert articulation errors. Within this context, we could also investigate whether different types of speech errors affect the ability of expert users to detect a synchronisation error, and whether there is a difference between typical and disordered speech.

In Section 6, our perceptual experiment revealed that experienced ultrasound users prefer the output of our system to hardware synchronisation. This indicates that we were able to improve synchronisation overall but does not tell us how good the automatic synchronisation was. Because rating and subjective scoring can be unreliable, ascertaining whether the automatic synchronisation was done to an acceptable level is best conducted in the context of a downstream task, as proposed above.

In Section 5, we trained our model on raw ultrasound data. However, other ultrasound systems used within the speech community produce DICOM sequences, or video recordings of ultrasound already in transformed format (AVI, MP4). Future work can explore transforming our data first and then training the model directly on the trans-

formed images to make it applicable to these other formats.

We can also extend our work to coronal ultrasound data. Because our model was trained on midsagittal utterances with the tip of the tongue to the right, we did not apply it to coronal Cleft utterances. In the future, we can explore collecting coronal images, validating their hardware synchronisation, and using them to adapt our model to this different orientation.

One limitation of our approach, which we identified while preparing the experiment in Section 6, is the need to specify the range of candidate offsets as input, by examining some samples of poorly synchronised data. This domain knowledge can restrict our ability to integrate the model into a data pre-processing pipeline. In the future, we will explore ways to eliminate the need to specify the range of offsets as input.

9. License and distribution

Acknowledgements

We thank the speech and language therapists, and speech scientists who took part in our perceptual experiments. We thank the children who took part in the "visualising speech" project, and their parents for allowing us to share the data with the research community. This work was funded by EPSRC Healthcare Partnerships Programme, grant number EP/P02338X/1 (Ultrax2020: http://www.ultrax-speech.org), and Action Medical Research, grant number GN2544 (Visualising Speech).

References

Ahn, S. (2018). The role of tongue position in laryngeal contrasts: An ultrasound study of english and brazilian portuguese. *Journal of Phonetics*, *71*, 451–467.

Articulate Instruments Ltd. (2010). *Articulate Assistant User Guide: Version 2.11*. Articulate Instruments Ltd. Edinburgh, United Kingdom. URL: http://www.articulateinstruments.com.

Bakst, S., & Lin, S. (2019). Post-collection ultrasound-audio synchronization. *The Journal of the Acoustical Society of America*, *146*, 3081–3081.

Bredin, H., & Chollet, G. (2007). Audiovisual speech synchrony measure: application to biometrics. *EURASIP Journal on Applied Signal Processing*, *2007*, 179–179.

Chen, S., Zheng, Y., Wu, C., Sheng, G., Roussel, P., & Denby, B. (2018). Direct, near real time animation of a 3d tongue model using non-invasive ultrasound images. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4994–4998).

Chen, Y., & Lin, H. (2011). Analysing tongue shape and movement in vowel production using ss anova in ultrasound imaging. In *Proceedings of International Congress of Phonetic Sciences (ICPhS)* (pp. 124–127). International Phonetic Association.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (pp. 539–546).

Chung, J. S., & Zisserman, A. (2016). Out of time: automated lip sync in the wild. In *Asian conference on computer vision* (pp. 251–263). Springer.

Chung, S., Chung, J. S., & Kang, H. (2019). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Cleland, J. (2018). Discussion regarding the use of hardware synchronisation by speech and language therapists. Personal Communication.

Cleland, J., Lloyd, S., Campbell, L., Crampin, L., Palo, J.-P., Sugden, E., Wrench, A., & Zharkova, N. (2020). The impact of real-time articulatory information on phonetic transcription: ultrasound-aided transcription in cleft lip and palate speech. *Folia Phoniatrica et Logopaedica*, *72*, 120–130.

Cleland, J., & Scobbie, J. M. (2021). The dorsal differentiation of velar from alveolar stops in typically developing children and children with persistent velar fronting. *Journal of Speech, Language, and Hearing Research*, (pp. 1–16).

Cleland, J., Scobbie, J. M., Heyde, C., Roxburgh, Z., & Wrench, A. A. (2017). Covert contrast and covert errors in persistent velar fronting. *Clinical Linguistics and Phonetics*, *31*, 35–55.

Cleland, J., Scobbie, J. M., Roxburgh, Z., Heyde, C., & Wrench, A. (2019). Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback. *Journal of Speech, Language, and Hearing Research*, *62*, 229–246.

Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A. (2017). DNN-based ultrasound-to-speech conversion for a silent speech interface. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, *120*, 407–415.

Denby, B., & Stone, M. (2004). Speech synthesis from real time ultrasound images of the tongue. In *Proceedings of ICASSP*. IEEE.

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J. M., & Wrench, A. A. (2018). UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Eshky, A., Ribeiro, M. S., Richmond, K., & Renals, S. (2019). Synchronising audio and ultrasound by learning cross-modal embeddings. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Fabre, D., Hueber, T., Bocquelet, F., & Badin, P. (2015). Tongue tracking in ultrasound images using eigentongue decomposition and arti-

[3] https://ultrasuite.github.io/data/cleft/

[4] https://github.com/aeshky/ultrasync

ficial neural networks. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Fabre, D., Hueber, T., Girin, L., Alameda-Pineda, X., & Badin, P. (2017). Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Communication*, *93*, 63–75.

Garau, G., Dielmann, A., & Bourlard, H. (2010). Audio-visual synchronisation for speaker diarisation. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, (pp. 113–121).

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. *Phonology and second language acquisition*, *36*, 315–328.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (pp. 1735–1742). volume 2.

Hensher, D. A., & Stopher, P. R. (1979). *Behavioural travel modelling*. Taylor & Francis.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, *52*, 288–300.

Hueber, T., Benaroya, E.-L., Denby, B., & Chollet, G. (2011). Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Hueber, T., Chollet, G., Denby, B., & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In *Proceedings of the International Seminar on Speech Production (ISSP)*.

ITU-R (1998). Recommendation ITU-R BT.1359: Relative timing of sound and vision for broadcasting.

Ji, Y., Liu, L., Wang, H., Liu, Z., Niu, Z., & Denby, B. (2018). Updating the silent speech challenge benchmark with deep learning. *Speech Communication*, *98*, 42–50.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems* (pp. 7763–7774).

Lawson, E., Stuart-Smith, J., Scobbie, J. M., Nakai, S., Beavan, D., Edmonds, F., Edmonds, I., Turk, A., Timmins, C., Beck, J. M. et al. (2015). Seeing speech: an articulatory web resource for the study of phonetics. URL: https://www.seeingspeech.ac.uk.

Lee-Kim, S.-I., Kawahara, S., & Lee, S. J. (2014). The 'whistled'fricative in xitsonga: Its articulation and acoustics. *Phonetica*, *71*, 50–81.

Mozaffari, M. H., Guan, S., Wen, S., Wang, N., & Lee, W. (2018). Guided learning of pronunciation by visualizing tongue articulation in ultrasound image sequences. In *Proceedings of International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1–5).

Porras, D., Sepúlveda-Sepúlveda, A., & Csapó, T. G. (2019). DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging. In *International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Ribeiro, M. S., Cleland, J., Eshky, A., Richmond, K., & Renals, S. (2021a). Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication*, *128*, 24–34.

Ribeiro, M. S., Eshky, A., Richmond, K., & Renals, S. (2019). Speaker-independent classification of phonetic segments from raw ultrasound in child speech. In *Proceedings of ICASSP* (pp. 1328–1332). IEEE.

Ribeiro, M. S., Sanger, J., Zhang, J.-X., Eshky, A., Wrench, A., Richmond, K., & Renals, S. (2021b). TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. In *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*. Shenzhen, China.

Richmond, K., Clark, R., & Fitt, S. (2010). On generating Combilex pronunciations via morphological analysis. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Richmond, K., Clark, R. A., & Fitt, S. (2009). Robust LTS rules with the Combilex speech technology lexicon. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Roxburgh, Z., Scobbie, J. M., & Cleland, J. (2015). Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback. In *Proceedings of International Congress of Phonetic Sciences (ICPhS)*. International Phonetic Association.

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, *9*, 1396–1403.

Spreafico, L., Pucher, M., & Matosova, A. (2018). Ultrafit: A speaker-friendly headset for ultrasound recordings in speech science. In *Proceedings of Interspeech*. International Speech Communication Association (ISCA).

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, *19*, 455–501.

Sugden, E., Lloyd, S., Lam, J., & Cleland, J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. *International Journal of Language & Communication Disorders*, *54*, 705–728.

Wilson, I., Gick, B., O'Brien, M., Shea, C., & Archibald, J. (2006). Ultrasound technology and second language acquisition research. In *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA)* (pp. 148–152).

Wrench, A. (2018a). Articulate Assistant Advanced (AAA): Research application for recording and analysing ultrasound, EPG, EMA and other instrumental data. Software v217.10.

Wrench, A. (2018b). Discussion regarding the hardware synchronisation method used in SonoSpeech and Articulate Assistant Advanced. Personal Communication.

Wrench, A. (2018c). SonoSpeech: Ultrasound application for recording, client assessment and visual feedback. Software v217.10.

Xu, K., Yang, Y., Stone, M., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel, P., & Denby, B. (2016). Robust contour tracking in ultrasound tongue image sequences. *Clinical linguistics & phonetics*, *30*, 313–327.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*, 23–43.

Zharkova, N. (2013). Using ultrasound to quantify tongue shape and movement characteristics. *The cleft palate-craniofacial journal*, *50*, 76–81.