Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors

Manuel Sam Ribeiro[a,*], Joanne Cleland[b], Aciel Eshky[a], 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

Speech sound disorders are a common communication impairment in childhood. Because speech disorders can negatively affect the lives and the development of children, clinical intervention is often recommended. To help with diagnosis and treatment, clinicians use instrumented methods such as spectrograms or ultrasound tongue imaging to analyse speech articulations. Analysis with these methods can be laborious for clinicians, therefore there is growing interest in its automation. In this paper, we investigate the contribution of ultrasound tongue imaging for the automatic detection of speech articulation errors. Our systems are trained on typically developing child speech and augmented with a database of adult speech using audio and ultrasound. Evaluation on typically developing speech indicates that pre-training on adult speech and jointly using ultrasound and audio gives the best results with an accuracy of 86.9%. To evaluate on disordered speech, we collect pronunciation scores from experienced speech and language therapists, focusing on cases of velar fronting and gliding of /r/. The scores show good inter-annotator agreement for velar fronting, but not for gliding errors. For automatic velar fronting error detection, the best results are obtained when jointly using ultrasound and audio. The best system correctly detects 86.6% of the errors identified by experienced clinicians. Out of all the segments identified as errors by the best system, 73.2% match errors identified by clinicians. Results on automatic gliding detection are harder to interpret due to poor inter-annotator agreement, but appear promising. Overall findings suggest that automatic detection of speech articulation errors has potential to be integrated into ultrasound intervention software for automatically quantifying progress during speech therapy.

Keywords: Speech sound disorders, Speech error detection, Ultrasound tongue imaging, Child speech

1. Introduction

Speech sound disorders (SSDs) are a common communication impairment in childhood (Wren et al., 2016). If left untreated, SSDs can have a negative impact on the social and emotional development of children and can lead to poor educational outcomes. For example, self-awareness of disordered speech contributes to low confidence in social situations when children engage with their peers or educators. In turn, this introduces communication barriers that lead to lower literacy levels (Johnson et al., 2010; Lewis et al., 2011; McCormack et al., 2011).

It is estimated that SSDs affect between 2.3% and 24.6% of children (Law et al., 2000; Wren et al., 2016). Speech and language therapy is often recommended, with the majority of interventions heavily reliant on auditory feedback. That is, the speech and language therapist (SLT) relies on their perceptual skills to give the child verbal feedback during intervention; and in turn the child relies on their perceptual skills to modify their articulations. This may also be accompanied by auditory cues describing where and how to place the articulators to produce the target sound. Interventions are often successful, especially for younger children (McLeod et al., 2020). However, some children do not respond well and the SSD becomes persistent. There is growing evidence that including visual biofeedback (VBF) during therapy is beneficial for such children (Sugden et al., 2019). VBF allows the visualization of the vocal tract during the speech production process, enabling children to view articulations in real-time.

With the widespread use of technology, there is increas-

*Corresponding author
Email addresses: sam.ribeiro@ed.ac.uk (Manuel Sam Ribeiro), joanne.cleland@strath.ac.uk (Joanne Cleland), aeshky@ed.ac.uk (Aciel Eshky), korin.richmond@ed.ac.uk (Korin Richmond), s.renals@ed.ac.uk (Steve Renals)

ing interest in automatically processing speech therapy tasks. This type of automation can be helpful to teachers and parents, who may use screening tools to determine the presence or absence of SSDs (Sadeghian & Zahorian, 2015; Ward et al., 2016); and to clinicians, who can save time on these time-consuming tasks (Ribeiro et al., 2019b). Clinicians and researchers are trained to use instrumented methods such as spectrograms or ultrasound tongue imaging to assist in the assessment, diagnosis, or quantification of treatment efficacy. These methods, however, can be laborious and impractical in the speech therapy clinic, as they still rely on manual annotation by the therapist or other trained professionals. Typical tasks include the identification of utterances spoken by the child, the identification of boundaries of target words, or measurements to determine correctness of speech articulations.

In this paper, we are concerned with the *automatic detection of speech articulation errors for speech therapy using ultrasound visual biofeedback*. This is intended as a tool for clinicians to automatically process data from ultrasound visual biofeedback assessment and therapy sessions. This work provides the following contributions: 1) the collection and analysis of pronunciation scores of velar fronting and gliding of /r/ given by experienced speech and language therapists; 2) a method for the automatic detection of speech articulation errors to be used by clinicians when processing data collected after therapy sessions; 3) an investigation of the impact of ultrasound tongue imaging for automatic error detection; and 4) an analysis of the impact of out-of-domain adult speech data for automatic error detection. Our method is evaluated on typically developing child speech and, more specifically, on cases of velar fronting and gliding of /r/ in Scottish English child speakers. These two errors are common in children with SSDs and amenable to intervention with ultrasound visual biofeedback (Sugden et al., 2019).

Section 2 provides background on speech disorders and recent evidence on the benefits of ultrasound visual biofeedback, as well as a review of recent literature on automatic speech articulation error detection. The data used throughout this work is described in Section 3. Section 4 describes a perceptual experiment where SLTs were asked to rate the goodness of pronunciation of phone instances, using ultrasound and audio data. Section 5 describes a set of experiments for the automatic scoring of phone pronunciations using ultrasound tongue imaging. Finally, Sections 6 and 7 provide an overall discussion and conclusion for this work, respectively.

Substitution	Description	Example
Fronting	Alveolars (/t,d/) replace velars (/k,g/)	*cookie → tootie*
Backing	Velars (/k,g/) replace alveolars (/t,d/)	*dog → gog*
Gliding	Glides (/w,j/ replace liquids (/r,l/)	*rabbit → wabbit*
Stopping	Stops (/p, d/) replace fricatives (/f,s/)	*zoo→ doo*
Labialisation	Labials (/p,b/) replace non-labials	*tie → pie*

Table 1: Selected examples of substitutions. A phone or group of phones is systematically replaced by another phone or group of phones.

2. Background

2.1. Speech sound disorders

SSDs occur when children exhibit difficulties in the production of speech sounds in their native language. *Organic speech sound disorders* denote difficulties that are associated with known causes. These causes may be motor or neurological (e.g. childhood dysarthria associated with cerebral palsy), structural (e.g. cleft lip and palate), or sensory (e.g. hearing impairments). *Functional speech sound disorders* are related to difficulties producing intelligible or acceptable speech with unknown causes. These disorders may be associated with motor production (articulation or motor speech disorders), or related to predictable or rule-based errors (phonological disorders) (ASHA, 2020).

Both types of SSDs can result in a variety of speech patterns. **Substitutions** are a common pattern where a phone or a group of phones is replaced by another phone or group of phones. Table 1 summarises common substitutions. We highlight here the two processes that are relevant to this work. *Fronting* occurs when phones that are produced towards back of the mouth (velars such as /k, g/) are replaced with phones produced towards the front of the mouth (alveolars such as /t, d/). This leads to word instances such as *cookie → tootie* or *gate → date*. *Gliding* occurs when liquids (e.g. /r, l/) are replaced with glides (e.g. /w/), originating cases such as *rabbit → wabbit* or *leg →weg*. Other examples of substitutions not listed in Table 1 are affrication, vowelization, depalatalization, or alveolarization (McLeod & Baker, 2017).

Beyond substitutions, we may observe **insertions** and **deletions** of phones in words (e.g. *black → buhlack*, *spoon → poon*). Alternatively, **assimilation** denotes cases when specific sounds are transformed due to the influence of those around it. For example, the process of nasal assimilation occurs when a non-nasal sound becomes nasal due to the presence of a nasal sound in the word (e.g. *bunny→ nunny*). Similarly, pre-vocalic voicing occurs when a voiceless phone becomes voiced when followed

by a vowel (e.g. *comb* → *gomb*). Additional phonological patterns may be influenced by **syllable structure**. For example, cluster reduction, where a consonant cluster is reduced (*plane*→ *pane*, *clean* → *keen*); the deletion of a consonant at the beginning (*bunny* → *unny*) or end of the syllable (*bus* → *bu*); or the deletion of weak or unstressed syllables in words (*banana* → *nana*). Other **phonetic distortions** may also be observed (for example, a lateral /s/).

Many of these processes are typical stages in the speech development of children. They are, however, expected to be eliminated as children reach a certain age. For instance, the typical age of elimination of velar fronting is around three years (McLeod & Crowe, 2018). On the other hand, because /r/ in particular is late acquired, gliding of this phone is typically eliminated around the age of five (McLeod & Crowe, 2018). Children that persist with one or more of these processes beyond their expected age of elimination usually require speech and language therapy.

2.2. Ultrasound visual biofeedback

In the context of speech sound disorders, visual biofeedback involves the use of instrumented methods to provide visual information regarding the position, movement, or shape of intra-oral articulators during speech production (Sugden et al., 2019). Common techniques to provide real-time visual biofeedback for speech therapy are electropalatography (EPG, Lee et al. (2009)), electromagnetic articulography (EMA, Katz et al. (2010)), and ultrasound tongue imaging (UTI, Sugden et al. (2019)). EPG uses an artificial palate to measure the contact points between the tongue and hard palate. However, the manufacture of custom-made palates incurs additional costs and may limit the use of this technique to a few patients. EMA requires the placement of sensor coils on the tongue and other articulators to measure their position over time, which can be both expensive and intrusive for children. Ultrasound tongue imaging uses diagnostic ultrasound operating in B-mode to visualise the tongue surface during the speech production process. A real-time B-mode ultrasound transducer is placed under the speaker's chin to generate a mid-saggital or coronal view of the tongue. This form of ultrasound is clinically safe, non-invasive, non-intrusive, portable, and relatively cheap (Stone, 2005). Figure 1 provides examples of ultrasound images of the tongue for a typically developing speaker.

Increasing evidence shows that ultrasound VBF can be beneficial for patients, therapists, and annotators (Bernhardt et al., 2005; Cleland et al., 2019, 2020). U-VBF is beneficial when used in intervention for a range of

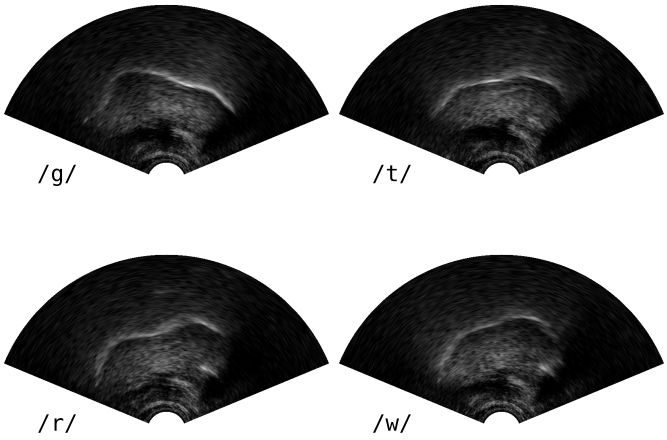

Figure 1: Ultrasound tongue images collected from a typically developing speaker (female, aged 11). Each frame is the mid-point of the phone showing a midsaggital view of the oral cavity with the tip of the tongue facing right. Samples extracted from the Ultrasuite repository (Eshky et al., 2018).

speech sound disorders, particularly if used in the initial stages of motor learning (Sugden et al., 2019). Related work suggests that U-VBF can be used as an objective measure of progress in intervention (Cleland & Scobbie, under revision), or to complement audio feedback and contribute to positive reinforcement (Roxburgh et al., 2015). U-VBF can also assist annotators in the identification of covert errors and increase inter-annotator agreement scores (Cleland et al., 2020). Additionally, U-VBF can contribute to the automatic processing of speech therapy recordings. Recent work used ultrasound data to develop tongue contour extractors (Fabre et al., 2015), animate a tongue model (Fabre et al., 2017), automatically synchronise therapy recordings (Eshky et al., 2019), and for speaker diarisation and alignment of therapy sessions (Ribeiro et al., 2019b). There are, however, several challenges associated with the automatic processing of ultrasound tongue images (Stone, 2005; Ribeiro et al., 2019a). Ultrasound output tends to be noisy, with unrelated high-contrast edges, speckle noise, or interruptions of the tongue surface. Image quality may also be affected by speaker characteristics (e.g. age and physiology) or session variability (e.g. incorrect or variable probe placement).

2.3. Automatic speech error detection

Automatic speech error detection aims to identify inaccurate productions of words or phones. These are often described in terms of insertions, deletions, and substitutions. Most studies adopt techniques from computer assisted pronunciation training, primarily developed for adult speakers using language learning systems (e.g. Witt

& Young (2000); Witt (2012); Hu et al. (2015)). This work is often considered part of the broader area of Computer Assisted Language Learning (CALL, Beatty (2013)). In children, speech error, or mispronunciation, detection can be used to assess reading levels (e.g. Black et al. (2010), Proença et al. (2018)), or with disordered speech for Computer Assisted Speech Therapy (CAST, e.g. Saz et al. (2009); Parnandi et al. (2015); Ahmed et al. (2018)). CALL systems are generally concerned with a global pronunciation score, which may or may not use the speaker's native language, while CAST systems aim to identify error types or underlying phonological processes.

Mispronunciation detection systems often use speech recognition techniques to compute pronunciation scores. Training data for the acoustic models is L2 speech for language learning, or typically developing speech for therapy applications. Other sources, if available, may consist of in-domain speech from the learners' native language or disordered speech. In-domain data can be used to develop extended search lattices accepting non-canonical pronunciation alternatives (Harrison et al., 2009; Ward et al., 2016; Dudy et al., 2018). The trained acoustic models and extended transducers then decode unseen utterances for which the text is known. To provide pronunciation scores, *likelihood-based systems* use the log-likelihoods generated by the models. The Goodness of Pronunciation score (GOP, Witt & Young (2000)) is a widely-used method for such systems. In its simplest form, the GOP score is the log-likelihood ratio between a target phone and a competing phone. Recently, Gaussian mixture models have been replaced by deep neural networks, with GOP-like scores defined over neural network posteriors (Hu et al., 2015). Alternatively, *classifier-based systems* use model outputs with supervised classifiers to determine error types or to provide more informed feedback. However, these methods require supervised in-domain data (e.g. phone-level annotated disordered speech), which are costly to acquire.

When annotated in-domain data is not available, a possible approach is to learn distributions over canonical training examples, such as typically developing speech. Unseen samples can then be compared against those distributions. Shahin et al. (2018) use one-class support vector machines to model the distribution of features describing manner and place of articulation. Wang et al. (2019) use Siamese recurrent networks, with positive and negative samples drawn from typically developing speech. In this paper, we adopt a similar strategy. Because we do not have annotated disordered speech for training, the acoustic model is trained only on typically developing speech. Scoring is based on a GOP-like score defined over neural network posteriors. Section 5 provides additional details on our model implementation.

3. Data

We use data from the **Ultrasuite repository**[1] (Eshky et al., 2018), consisting of synchronised ultrasound and audio data from child speech therapy sessions. Ultrasuite currently contains three datasets of child speech. Ultrax Typically Developing (UXTD) includes recordings of 58 typically developing children. The remaining datasets include recordings from children with speech sound disorders collected over the course of assessment and therapy sessions: Ultrax Speech Sound Disorders (UXSSD, 8 children) and Ultraphonix (UPX, 20 children). Assessment sessions denote recordings at various stages of therapy: baseline (before therapy), mid-therapy, post-therapy (immediately after therapy), and maintenance (several months after therapy). For the child speech datasets, ultrasound was recorded with an Ultrasonix SonixRP machine using Articulate Assistant Advanced (AAA, Articulate Instruments Ltd. (2010)) software at ~120fps with a 135° field of view. A single B-Mode ultrasound frame has 412 echo returns for each of 63 scan lines, giving a 63×412 "raw" ultrasound frame capturing a mid-sagittal view of the tongue. Samples from this data are illustrated in Figure 1.

To complement the Ultrasuite repository, we use the **Tongue and Lips corpus**[2](TaL, Ribeiro et al. (2021)). TaL is a corpus of synchronised ultrasound, audio, and lip videos from 82 adult native speakers of English. Ultrasound in the TaL corpus was recorded using Articulate Instruments' Micro system (Articulate Instruments Ltd., 2010) at ~80fps with a 92° field of view. TaL used a different transducer than the one in the Ultrasuite data collection. Because of this, an ultrasound frame of the TaL corpus contains 842 echo returns for each of 64 scan lines (64×842 "raw" ultrasound frame).

4. Expert speech error detection

In this section, our **goal** is the collection of pronunciation scores for speech segments produced by children with speech sound disorders. This data is to be used in the evaluation of the automatic error detection systems described in Section 5. We recruited Speech and Language Therapists with experience using ultrasound visual biofeedback

[1] https://www.ultrax-speech.org/ultrasuite

[2] Available via the Ultrasuite Repository, see footnote 1.

and who routinely work with Scottish English-speaking children. The collection of these scores aimed to simulate the process therapists undergo after collecting data from speech therapy sessions. We are interested in the processes of velar fronting and gliding of /r/, therefore we focused on productions of /k, g/ (*velars*) and pre-vocalic /r/ (*rhotic*). We expected SLTs to be able to identify correct and incorrect productions of velars and rhotics with good reliability.

4.1. Data preparation

We used data from the eight children available in Ultrasuite's UXSSD dataset. Children in this dataset were treated for velar fronting, therefore we expected to observe an increasing number of correct velar productions throughout assessment sessions. Children's response to intervention is reported in Cleland et al. (2015). Because intervention for these children did not focus on correcting rhotic productions, this led to an imbalanced set of correct and incorrect rhotic samples. We pre-selected words containing the target velar and rhotic phones and occurring within prompts of type "A" (single words) in assessment sessions (baseline, mid-therapy, post-therapy, and maintenance). We discarded words that contained more than one instance of a target phone (e.g. "cake") or corrupted word instances (e.g. overlapping or unintelligible speech, other background noise, etc). From the pre-selected word list, we randomly sampled 96 word instances per child. Samples were balanced across assessment sessions and across velars and rhotics. For each child, an assessment session contained 24 word instances (12 velars and 12 rhotics). Where possible, samples were also balanced for the position of the target phone in the word (initial, medial, or final). The final set of samples consisted of a total of 768 word instances with a vocabulary of 148 words, which we denote as the **main set**.

We generated a set of additional samples from one of the speakers available in Ultrasuite's UPX dataset. We selected speaker 04M, treated for velar fronting and reported to have good improvement after intervention (Cleland et al., 2019). As the speakers in the main set, this speaker was also not treated for gliding. The sampling process was repeated and a total of 24 word instances were selected (12 velars and 12 rhotics), balanced across assessment sessions. We denote this set of samples the **control set**.

4.2. Method

To annotate the word instances, we recruited 8 annotators. All annotators were SLTs with more than 4 years of experience and who routinely work with children speaking Scottish English. Additionally, the annotators had at least 3 years of experience with ultrasound visual biofeedback, with two annotators having more than 10 years of experience. Each annotator was assigned 96 words from the main set and the 24 words from the control set. Words taken from the main set were selected such that they were produced by a single child. Therefore, each SLT annotated data from two children (one main and one control). For intra-annotator agreement, 20% of the words (24 samples) were repeated in the annotation list. Of these, 12 were taken from the control set and 12 from the main set. Each SLT annotated a total of 144 word instances.

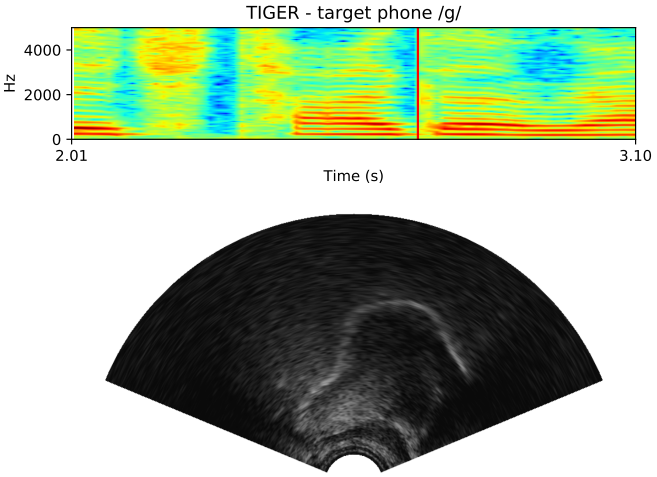

Figure 2: Video frame for the word "tiger" produced by speaker 02M in a post-therapy assessment session. The annotator is requested to score the target phone /g/ in a word medial position. The vertical red line in the spectrogram indicates the current position of the video.

Results were collected via a web interface displaying a video of each word sample separately. The video contained the spectrogram, ultrasound images of the tongue, and the audio for each word. Figure 2 illustrates one video frame extracted from one of the samples. Annotators were allowed to play videos at normal, half, or quarter speed up to a maximum of 6 total playbacks.

Annotators were instructed to rate the target phone on a 5-point Likert scale, where 1 indicates *wrong pronunciation* and 5 indicates *perfect pronunciation*. The first question requested a score for the target phone (e.g. "Please rate the velar /k/ in the sample"). This score is denoted the **primary score**. If the annotator scored 3 or lower in the primary score, we requested a **secondary score**. The secondary score asked the annotator to rate the target phone with respect to an expected substitution. (e.g. "Please rate the target phone for alveolar substitution" or "Please rate

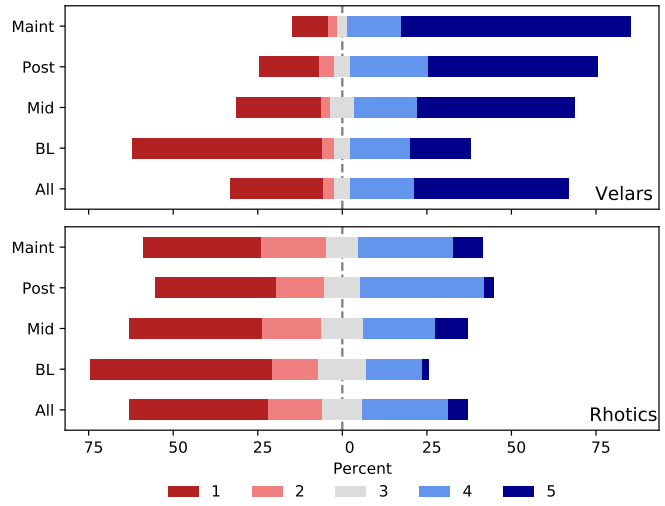

Figure 3: Normalised distribution of primary scores for Velars and Rhotics, ordered chronologically (bottom up) by assessment session: baseline (BL), mid-therapy (Mid), post-therapy (Post), and maintenance (Maint).

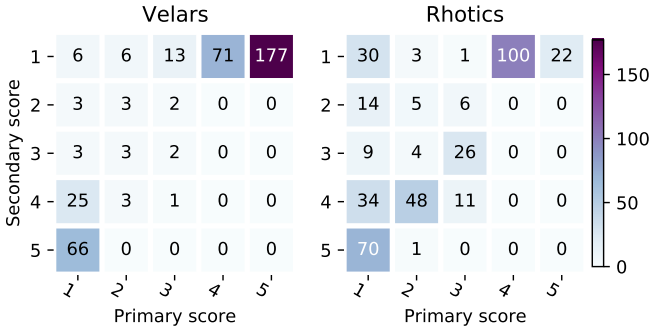

Figure 4: Frequency of primary and secondary scores given by annotators for Velars and Rhotics in the main set with duplicate entries removed, where 1 indicates wrong pronunciation and 5 indicates perfect pronunciation.

the target phone for gliding substitution"). An optional field allowed annotators to provide a short comment for each sample.

Given the primary score s_p and the secondary score s_s, we determine a **combined score** s_c defined as $s_c = \log(s_p) - \log(s_s)$. Because we did not request a secondary score when perfect pronunciation was rated for the target phone, we assumed a value of 1 for s_s when computing the combined score. The score s_c is positive if there is a preference for the primary class (e.g. velars or rhotics) and negative if there is a preference for the secondary class (e.g. alveolars or glide). If the annotator gave the same primary and secondary scores, then this uncertainty is represented in s_c with 0. This preference can be further simplified to produce a **binary score** s_b. For each sample, positive values of the combined scores are treated as correct productions of the primary class and negative or zero values as incorrect productions.

4.3. Results

Figure 3 shows the normalised distribution of the primary score for all annotated samples, ordered chronologically by session. The number of incorrect velars across assessment sessions decreases over time, while the distribution for rhotics is more or less stable. This is expected since intervention for these children focused on velar fronting and production of /r/ was not addressed.

Figure 4 shows the frequency of primary and secondary scores for velars and rhotics in the main set. We remove duplicate samples used for intra-annotator agreement, keeping the score of the first sample to be rated. For this work, we are primarily interested in the correct production of velars and rhotics and clear cases of substitutions (fronting and gliding). Cases of correct pronunciations for the expected class are identified by a high primary score (4 or 5). Cases of velar fronting or gliding are identified by a low primary score (1 or 2) and a high secondary score (4 or 5). We observe from Figure 4 that 342 out of the 384 velar samples (89.06%) fall under one of these two cases. Of these samples, 248 are correct velars and 94 are alveolar substitutions. Rhotics include a smaller number of correct productions or gliding (275 out of 384 samples, 71.61%). Of these, 122 are marked as a correct production of /r/, while 153 denote cases of gliding.

Some annotators used the optional comment field to elaborate on their score, particularly for incorrect cases that were not instances of velar fronting or gliding. For velars, some of the cases were reported to be uvular, palatal, or postalveolar realisations, or omitted phones. For rhotics, most of the non-typical scores indicated deletion of /r/ with some cases reporting a distortion towards a labiodental approximant.

4.4. Annotator agreement

We compute inter-annotator agreement using the control set of samples, rated by all annotators. Duplicates are removed by choosing the first rating and discarding the second. Intra-annotator agreement is computed on the 20% duplicate samples, half from the main set and half from the control set.

To measure global agreement, we use *Krippendorf's* α (Krippendorff, 2004), which computes annotator agreement for multiple annotators and supports several levels

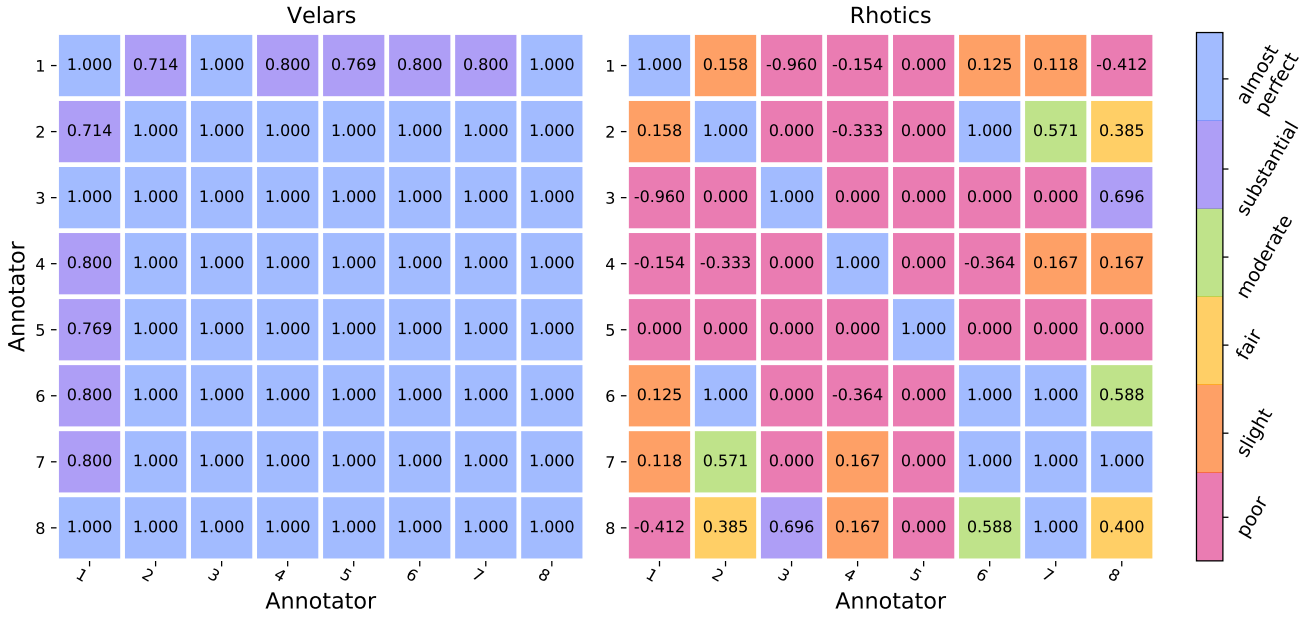

Figure 5: Pairwise Cohen's κ for the binary score, excluding cases not rated as correct productions or clear substitutions. Diagonal values show κ for intra-annotator agreement, while off-diagonal shows pairwise inter-annotator agreement. Each cell is colour-coded according to the levels of agreement proposed by Landis & Koch (1977, pp 164-165).

	Primary	**Combined**	**Binary**
All	0.601	0.578	0.579
Velars	0.883	0.868	0.946
Rhotics	0.210	0.117	0.050

Table 2: Krippendorf's α for primary score s_p, combined score s_c, and binary score s_b computed for all annotators across all, velar, and rhotic samples. Agreement for binary score excludes samples not rated as correct productions or clear substitutions.

of measurements. We compute α using a difference function for ordinal data for the primary score, a function for interval data for the combined score, and a function for nominal data for the binary score (Krippendorff, 2011). According to Krippendorff (2004, pp 241-243), $\alpha > 0.8$ indicates reliable data, while $0.667 \leqslant \alpha \leqslant 0.8$ indicates moderately reliable data. When $\alpha < 0.667$, the data should be considered unreliable. Table 2 shows α values for primary, combined, and binary scores. For the binary score, we exclude samples not rated as correct productions or clear substitutions. Overall annotator agreement appears to be very good for velar samples and poor for rhotic samples.

We measure pairwise agreement using Cohen's κ (Cohen, 1960), which measures agreement between two raters on categorical data. The *kappa* statistic is a standardised metric where $\kappa \in [-1, 1]$, with $\kappa = 0$ denoting chance agreement and $\kappa = 1$ denoting perfect agreement. Traditionally, Cohen's κ is discussed according to the five agreement levels suggested by Landis & Koch (1977, pp 164-165). These group values of κ into: poor ($\kappa \leq 0.0$), slight ($0.0 < \kappa \leqslant 0.2$), fair ($0.2 < \kappa \leqslant 0.40$), moderate ($0.40 < \kappa \leqslant 0.60$), substantial ($0.60 < \kappa \leqslant 0.80$), and almost perfect ($0.80 < \kappa \leqslant 1.0$) agreement. Figure 5 visualises pairwise annotator agreement for the binary score. Off-diagonal values denote pairwise inter-annotator agreement on the control set, whereas diagonal values denote intra-annotator agreement on the duplicate samples.

According to Figure 5, scores provided for the velar samples are very consistent and reliable, with perfect agreement across most raters. This is observed for inter and intra-annotator scores. Annotator 1 has substantial agreement with some of the other raters, but not perfect. Excluding samples rated by annotator 1 leads to improved global inter-annotator agreement for velar samples on the combined score ($\alpha = 0.922$). Results for the rhotic samples, however, indicate a substantial disagreement between annotators. We observe a perfect agreement for intra-annotator scores across all annotators except annotator 8. Rhotic agreement between annotators 6, 7, and 8 is higher than other raters, with perfect or moderate agreement. However, considering only those three raters, global agreement on the combined score is still lower than the moderate reliability threshold for Krippendorf's α ($\alpha = 0.631$). There are various reasons that

Data set	Source	Speakers	Samples	Notes
Train	UXTD	45	8302	Child in-domain train data
Train	TaL	81	81193	Adult out-of-domain train data
Validation	UXTD	5	534	Child in-domain validation data
Test	UXTD	13	901	Typically developing evaluation set
Test	UXSSD	8	768	Disordered speech evaluation set

Table 3: Datasets used for automatic speech error detection. All sets contain samples from all places of articulation, except the disordered speech evaluation set, which contains the velar (384) and rhotic (384) samples rated by annotators in Section 4.

could explain the agreement discrepancy between velar and rhotic samples. We provide further insights into these results in Section 6. However, these results indicate that we should use rhotic scores carefully when evaluating automatic error detection systems in the next section.

5. Automatic speech error detection

In this section we investigate the automatic detection of speech errors in typically developing and disordered Scottish English child speech. The proposed system is designed as a tool to be used by Speech and Language Therapists on data collected from speech therapy and assessment sessions. Therefore, we evaluate model scores with the expert scores for velar fronting and gliding of /r/ provided by therapists in Section 4. Our **goal** is to investigate the proposed system's ability to simulate expert behaviour in the detection of substitution errors. Additionally, we aim to analyse the contribution of *ultrasound tongue imaging* and *out-of-domain adult data* on the automatic speech error detection.

5.1. Data preparation

For **training data**, we use Ultrax Typically Developing (UXTD), which collected data from 58 child speakers. UXTD includes a subset of utterances with manually-annotated word boundaries for 13 speakers. We save data from those speakers for evaluation. From the remaining 45 speakers, we randomly select 40 speakers for training and 5 speakers for validation. The TaL corpus of adult speech is used as an additional source of training data. We use the TaL80 dataset, containing data from 81 speakers.

For **evaluation data**, we use an evaluation set of disordered speech samples and an evaluation set of typically developing speech samples. The disordered speech samples consist of the *main set* rated by expert SLTs, described in Section 4. This is a set of 768 word instances from the UXSSD dataset. The typically developing samples are extracted from the UXTD dataset, which includes 220 utterances with manually annotated word boundaries.

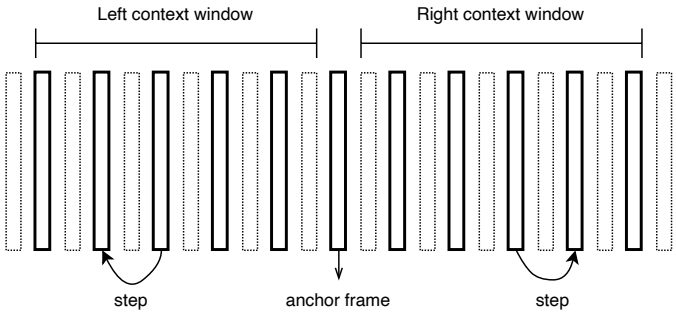

Figure 6: Sample build process. A sample is constructed around an anchor frame, typically the mid-point of a phone instance. Context windows are fixed at 100ms. Step is set such that 5 context frames are extracted for audio and 4 context frames for ultrasound.

These utterances are produced by 13 speakers, disjoint from those in the training and validation sets. After pre-processing, the typically developing evaluation set consists of 901 phone instances extracted from 866 words with a vocabulary of 153 words.

Because output classes are unbalanced, we control the number of samples per class for the training data. For classes that are under-represented, we retrieve additional examples. This is done by perturbing the anchor frame by up to 40ms for under-represented classes. For classes that are over-represented, we randomly sample 1000 and 10000 examples for the UXTD and TaL sets, respectively. After balancing and pre-processing, we have a total of 8302 for the UXTD training data. The TaL corpus is larger than UXTD and has a total of 81193 samples. Table 3 shows the datasets used in this section, and their respective number of speakers and number of samples.

The Kaldi speech recognition toolkit (Povey et al., 2011) is used to force-align all datasets at the phone level using the reference audio. For the evaluation data, we constrain the phone alignment to the manually verified word boundaries. The phone set is a Scottish accent variant of the Combilex lexicon (Richmond et al., 2010, 2009). We discard silence segments and vowels from the phone set and map the remaining phones onto one of nine classes corresponding to place of articulation: *alveolar, dental, labial, labiovelar, lateral, palatal, postalveolar, rhotic*, and *velar*. From the training data, we exclude phone instances that do not have parallel audio and ultrasound. These instances occurred when audio started recording before the ultrasound. For the UXTD data, there were 91 segments excluded due to early start.

We use Kaldi to extract Mel-frequency cepstral coefficients (MFCCs) for the audio signal. MFCCs are commonly used for speech recognition, with good results re-

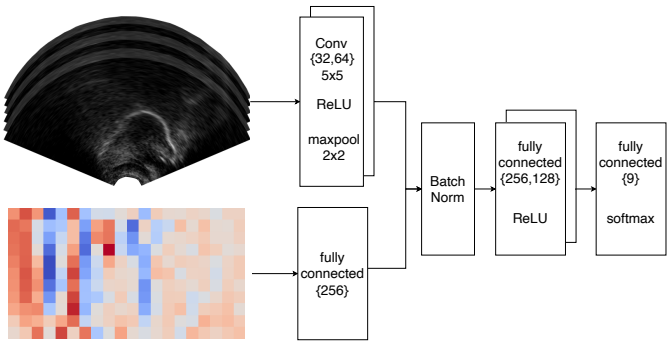

Figure 7: Convolutional neural network architecture for classifier using ultrasound and audio (MFCCs) inputs.

ported for child speech recognition (Shivakumar et al., 2014). Waveforms are downsampled to 16KHz and features computed every 10ms over 25ms windows. We keep 20 cepstral coefficients and append their respective first and second derivatives for a total of 60 features. A high number of cepstral coefficients is helpful for child speech processing (Li & Russell, 2001). Ultrasound frames are individually reshaped to 63×103 using bi-linear interpolation. A single sample consists of an anchor frame and a set of context frames. The anchor frame is fixed at the mid-point of each phone instance and the set of context frames are extracted over a fixed sized window of 100 ms to the left and right of the anchor frame. Because of the different frame rates, the number of frames in the context window is different for the ultrasound and audio streams. For the audio, each context window corresponds to 10 frames. For ultrasound, the context window corresponds to 12 frames for Ultrasuite data and to 8 frames for TaL data. Over each context window, we extract 5 MFCC frames and 4 ultrasound frames, with the step size set separately to account for the respective frame rates. Figure 6 illustrates the sample build process.

5.2. Model architecture and training

The adopted model architecture, illustrated in Figure 7, largely follows that of earlier work (Ribeiro et al., 2019a,b). The ultrasound stream is processed by two convolutional layers. These layers use 5×5 kernels with 32 and 64 filters, respectively, and ReLU activation functions. Each convolutional layer is followed by max-pooling with a 2×2 kernel. The sequence of frames for the audio stream is flattened and processed by a fully-connected layer with rectified linear units. When using the ultrasound and audio streams, the features are concatenated at this stage. The batch normalized features are then processed by two fully-connected layers with ReLU activation functions and an output fully-connected layer followed by the softmax function.

Models are optimized via Stochastic Gradient Descent with minibatches of 128 samples and an L2 regularizer with weight 0.1. We train models on the UXTD data or on the pooled TaL and UXTD data. When using the UXTD training data, systems are optimized for 200 epochs with a learning rate of 0.1. With the pooled dataset, systems are optimized for 50 epochs with an identical learning rate of 0.1. After each epoch, the model is evaluated on the validation data and we keep the best model across all epochs. We fine-tune systems trained on the pooled data on the UXTD data. Models that are fine-tuned reduce the learning rate to 0.001 and are optimized for 100 epochs.

5.3. Scoring

The output of the classifier is a probability distribution over the nine places of articulation. To score an input phone instance x, we consider an expected class y and a competing class c. The model score s_m is then computed as

$$s_m = \log(p(y|x)) - \log(p(c|x)) \tag{1}$$

The expected class may be the canonical phone class, such as a velar or a rhotic. The competing class is a possible substitution, such as an alveolar or a labiovelar approximant. If no competing class is given, we can estimate it and compute the phone score with

$$c = \arg\max_{q \in Q - \{y\}} p(q|x) \tag{2}$$

where Q is the set of places of articulation considered by the model. This method is related to the Goodness of Pronunciation score (Witt & Young, 2000; Hu et al., 2015). As with the combined expert score s_c (Section 4.2), the magnitude of the model score encodes certainty, whereas the sign encodes preference. A positive score indicates preference for the expected class, while a negative score indicates preference for the competing class. We simplify model and combined expert scores onto a binary correct/incorrect label $b(s)$ for error detection according to:

$$b(s) = \begin{cases} 0 & \text{if } s > k \\ 1 & \text{if } s \leq k \end{cases} \tag{3}$$

where s is either s_c or s_m and k is a configurable threshold. Unless otherwise stated, results presented in this work use $k = 0$, which treats uncertainty in the model score ($s_m = 0$) as an error. For the purposes of this analysis, uncertainty is not applicable to the combined expert score because we retain only cases that are correct or clear substitutions.

Training Data	Alveolar	Dental	Labial	Labiovelar	Lateral	Palatal	Postalveolar	Rhotic	Velar	Global
					Audio					
UXTD	72.36%	46.77%	64.00%	52.5%	84.85%	64.52%	68.66%	85.44%	75.62%	70.81%
Joint	70.56%	35.14%	67.21%	52.11%	74.02%	50.00%	71.64%	88.64%	70.25%	65.93%
Joint (+fine-tuning)	75.49%	39.73%	66.67%	59.42%	80.56%	66.67%	70.59%	85.71%	77.65%	72.03%
					Ultrasound					
UXTD	78.32%	53.62%	59.21%	74.42%	78.95%	25.93%	50.67%	67.57%	88.27%	69.48%
Joint	79.38%	60.34%	47.11%	67.35%	84.38%	22.22%	41.00%	73.13%	92.52%	68.37%
Joint (+fine-tuning)	84.05%	61.11%	60.82%	76.00%	84.91%	38.89%	48.81%	79.26%	94.89%	76.14%
					Audio+Ultrasound					
UXTD	80.10%	81.08%	83.87%	82.61%	90.18%	65.52%	58.33%	74.83%	89.58%	80.47%
Joint	83.25%	76.74%	83.33%	63.08%	73.19%	66.67%	76.19%	91.59%	93.45%	81.80%
Joint (+fine-tuning)	87.94%	76.47%	87.78%	72.31%	91.82%	58.82%	75.38%	94.34%	95.58%	**86.90%**
Number of samples	196	37	62	46	112	29	84	143	192	901

Table 4: Accuracy for the typically developing evaluation set across all places of articulation. Global accuracy is an average of all places of articulation, weighted by the number of samples for each class. The Joint training data denotes the pooled UXTD and TaL data. Highlighted results in bold indicate best overall performance.

Training data	Precision	Recall	F1-Score	Accuracy
		Audio		
UXTD	0.384	0.524	0.443	64.1%
Joint	0.417	0.585	0.487	66.5%
Joint (+fine-tuning)	0.393	0.537	0.454	64.9%
		Ultrasound		
UXTD	0.670	0.842	0.746	84.4%
Joint	0.702	0.805	0.750	85.4%
Joint (+fine-tuning)	0.677	0.768	0.720	83.7%
		Audio+Ultrasound		
UXTD	0.732	0.866	**0.793**	**87.7%**
Joint	0.704	0.695	0.699	83.7%
Joint (+fine-tuning)	0.681	0.756	0.717	83.7%

Table 5: Results for velar fronting error detection using the UXSSD velar samples rated by all annotators, except annotator 1. Scores are computed using "velar" and "alveolar" as expected and competing classes respectively.

5.4. Results

We evaluate model performance on the **typically developing** set. Table 4 shows accuracy results, which are computed across examples of all output classes. Systems trained on the joint UXTD and TaL data underperform when compared with systems trained only on the UXTD data, even though there is more training data available. However, fine-tuning the pre-trained joint model on the UXTD data leads to the best performance.

Comparing systems using only one modality, accuracy results are better for ultrasound when using additional TaL data. As expected, systems using both audio and ultrasound provide the best results. Observing accuracy separately for each class, we observe that *labial*, *palatal*, *postalveolar*, or *rhotic* speech sounds have better results when using only audio compared to using only ultrasound. The remaining speech sounds have better results with ultrasound tongue imaging. Such differences are expected due to the individual characteristics of speech sounds. For example, labial sounds do not rely on tongue movement, so they are not expected to benefit much from ultrasound tongue imaging alone. On the other hand, velar and alveolar sounds have well-defined tongue shapes on the midsaggital plane, so we would expect ultrasound data to be the primary contributor when identifying them. We also observe that accuracy improves across all classes when using both modalities as input. These results meet our expectations that ultrasound and audio complement each other well and that additional out-of-domain training data is beneficial. Similar findings were reported on related tasks, such as speaker diarisation and word alignment of speech therapy sessions (Ribeiro et al., 2019b).

Table 5 shows results for **velar fronting error detection**. These are computed over samples identified by annotators as correct velar productions or alveolar substitutions (see Figure 4). We exclude samples rated by annotator 1, due to less than perfect agreement with other annotators. Results are computed on $b(s)$ using the combined expert score s_c and the model score s_m with an expected *velar* class and a competing *alveolar* class.

We observe that ultrasound is more suited than audio to discriminate between velar and alveolar productions, although systems using both data streams have the best results. There are no performance improvements to the systems using the joint dataset and fine-tuning when compared to the system using only typically developing child data. This is an interesting observation, as results on the

typically developing dataset indicate that using additional training data and fine-tuning is beneficial. On the typically developing data, the individual accuracy for the velar and alveolar classes increases with more data and training. Considering the system using both audio and ultrasound and comparing the UXTD and fine-tuned systems, velar accuracy increases from 89.58% to 95.58% and alveolar accuracy increases from 80.10% to 87.94%. The discrepancy observed between the typically developing set and velar fronting error detection could be attributed to challenges associated with speaker's data. Speaker performance can vary substantially, particularly when using ultrasound data (Ribeiro et al., 2019a). This observation can be further supported by measuring agreement between model and expert binary scores for each of the eight speakers. Using Cohen's κ (Cohen, 1960), models and expert scores have near perfect agreement for speakers 1 and 3 ($\kappa > 0.8$), substantial agreement for speakers 4 and 7 ($0.6 < \kappa \leq 0.8$), moderate agreement for speakers 5 and 6 ($0.4 < \kappa \leq 0.6$), and no or slight agreement ($\kappa \leq 0.2$) for speakers 2 and 8.

In Section 4.4, we reported intra- and inter-annotator agreement for the scoring of rhotic productions. Unlike velars, expert scores for rhotics were shown to have very low inter-annotator agreement. For this reason, results for **gliding error detection** should be interpreted carefully. However, intra-annotator agreement was reliable and consistent for all annotators except annotator 8 (Figure 5). Therefore, we opt to analyse the results for gliding error detection separately for each speaker.

We note from Table 4 that classification accuracy for rhotics and labiovelars is good across all classifiers on the typically developing evaluation set. We observe that classification of rhotic instances benefits more from audio (85.71%) than ultrasound (79.26%). The labiovelar class, on the other hand, achieves higher accuracy when using only ultrasound (76.0%) than when using only audio (59.42%). Jointly using audio and ultrasound improves accuracy for rhotics (94.34%) but not for labiovelars (72.31%). The average accuracy for rhotic and labiovelars when using ultrasound and audio is 78.72% when training only on the UXTD data. This accuracy slightly decreases when jointly training on the TaL corpus (77.34%), but improves when fine-tuning on the UXTD data (83.33%). These results, however, relate to the typically developing evaluation set. As observed with the velar case, they may not transfer in the same way to error detection. Nevertheless, we select the fine-tuned system using both ultrasound and audio to analyse speaker-wise results for gliding error detection.

Speaker	N	Precision	Recall	F1-Score	Accuracy	Cohen's κ
1	41	0.778	0.467	0.583	75.6%	0.426
2	29	1.000	0.071	0.133	55.2%	0.074
3	36	0.900	0.474	0.621	69.4%	0.404
4	11	0.833	1.000	0.909	90.9%	0.820
5	42	0.964	0.675	0.794	66.7%	0.045
6	36	1.000	0.639	0.780	63.9%	0.000
7	45	0.500	1.000	0.667	97.8%	0.656
8	35	0.692	0.783	0.735	62.9%	0.123

Table 6: Speaker-wise results for gliding error detection using the UXSSD rhotic samples identified as correct productions or gliding substitutions. Cohen's κ is calculated on expert and model binary scores. Results are generated by the fine-tuned system using both audio and ultrasound, with scores computed using "rhotic" and "labiovelar" as expected and competing classes, respectively.

Table 6 shows speaker-wise results for gliding error detection and Figure 8 shows their respective confusion matrices. We observe that the scores given by the expert annotators can vary per speaker. For example, most samples produced by speakers 5 and 6 were marked as gliding cases by their respective annotators. The limited number of correct productions influences the calculation of Cohen's κ, leading to poor agreement even though accuracy and F_1 are high. On the other hand, most samples by speaker 7 were marked as correct instances. The classifier appears to behave similarly with data from speakers 2 and 7, with most samples classified as correct rhotic instances. This behaviour is in agreement with the expert for speaker 7, but not for speaker 2. Most errors produced by the model are Type II errors (false negatives). This might be due to the lower performance of the competing labiovelar class, as observed on the typically developing set. Because the classifier is more confident when scoring rhotics, there is a limited number of Type I errors (false positives). Due to the low inter-annotator agreement, it is not clear whether these differences are due to the scores provided by the annotators or due to challenges associated with speaker or recording variability.

6. Discussion

Considering the scores provided by the expert Speech and Language Therapists, we observed that results for velar samples are consistent and reliable, meeting our initial expectations. The high inter-annotator agreement, excluding rater 1, for the combined score ($\alpha = 0.922$) suggests that the data provided by the experienced SLTs can be used for the evaluation of automatic methods. The scores provided on the rhotic samples, however, were less consistent and did not meet our original expectations. There are various reasons why this might have occurred. Be-

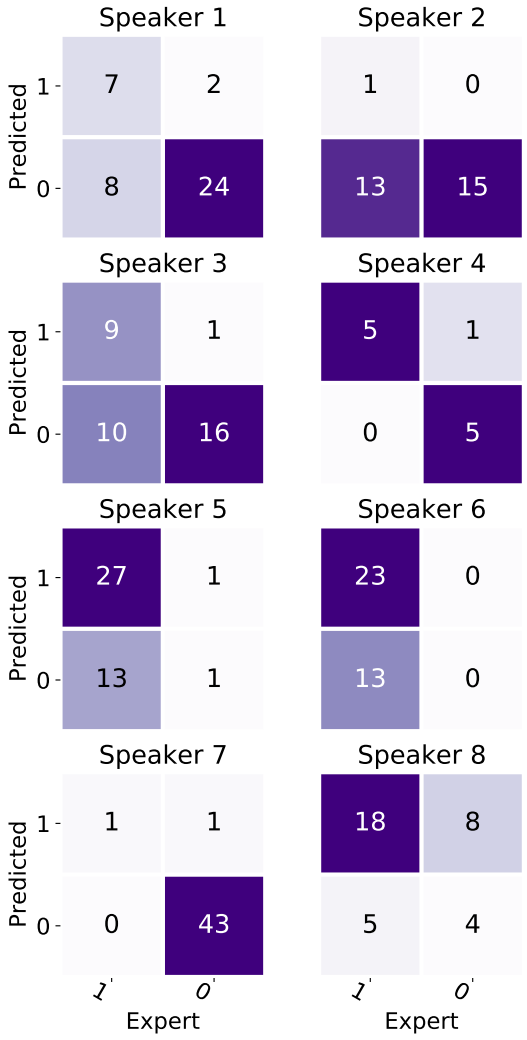

Figure 8: Confusion matrices for gliding error detection produced by the fine-tuned system using audio and ultrasound for all 8 speakers in the UXSSD evaluation set. Correct rhotic productions are denoted by 0 and gliding substitutions are denoted by 1.

cause children from the Ultrasuite repository were not treated for the production of rhotics, correct and incorrect samples were unbalanced in the annotation list. This may have affected the judgements made by the annotators, who expected to encounter some correct productions of /r/. Including samples from typically developing children could have mitigated this issue and helped anchor the scale for correct productions. Additionally, /r/ is a less common target for intervention in the UK (Wren et al., 2016). This might have affected the behaviour of the experienced SLTs, who, although able to discriminate between correct and incorrect productions, have less experience when evaluating this particular speech sound clinically. Related to this observation, there is a wide range of socially acceptable productions of /r/ in the United Kingdom (Scobbie, 2006; Lawson et al., 2011), which motivated the choice of gliding for analysis in this work. Unlike velars, which have a relatively consistent tongue shape between speakers, rhotics can be produced with a wide variety of tongue shapes (Boyce, 2015), potentially making it more difficult to judge acceptability using ultrasound. To account for the wide range of acceptable productions of /r/, annotators could browse through a set of samples drawn from typically developing children.

With respect to automatic scoring of speech articulation errors, our results indicate that expert behaviour can be simulated with an acceptable level of accuracy for velar fronting error detection. The best performing system correctly detected 86.6% of all errors identified by SLTs. Out of all the segments identified as errors, 73.2% of those are correct. When evaluating systems, we assumed a threshold $k = 0$ to compute the final score according to Equation 3. The threshold k is a parameter configured by the user, allowing some control over precision and recall. Figure 9 illustrates model scores and the impact of k on precision and recall. As expected, most of the uncertainty with respect to the true label occurs around $k = 0$. For the system in Figure 9, the F_1 score improves from 0.793 to 0.817 when $k = -0.4$.

Even though changing k might result in slight improvements, the ranking of the systems across all conditions remains the same. Ultrasound tongue imaging has a positive contribution to the overall accuracy of the models, when used by itself or together with audio features. Considering out-of-domain data, results show that model performance can be improved when pre-trained on adult speech data. Overall performance increases further when fine-tuning models on in-domain data. This is observed when evaluating on typically developing speech, but not when detecting errors on disordered speech data. This discrepancy could be caused by differences in the two datasets. The typically developing set (UXTD) and disordered speech set (UXSSD) were collected separately, with different purposes and conditions (Eshky et al., 2018). This might lead to domain mismatches between training (UXTD) and test (UXSSD) data. Although the model achieves better accuracy on typically developing data, it may not generalise to the different disordered speech domain. Differences between the two datasets include speaker characteristics, ultrasound probe placement, or acoustic variability due to room conditions or hardware used for data collection.

Results for gliding error detection are harder to interpret due to low inter-annotator agreement. We do observe reasonable accuracy for some speakers in the evaluation set. However, further work should investigate primarily

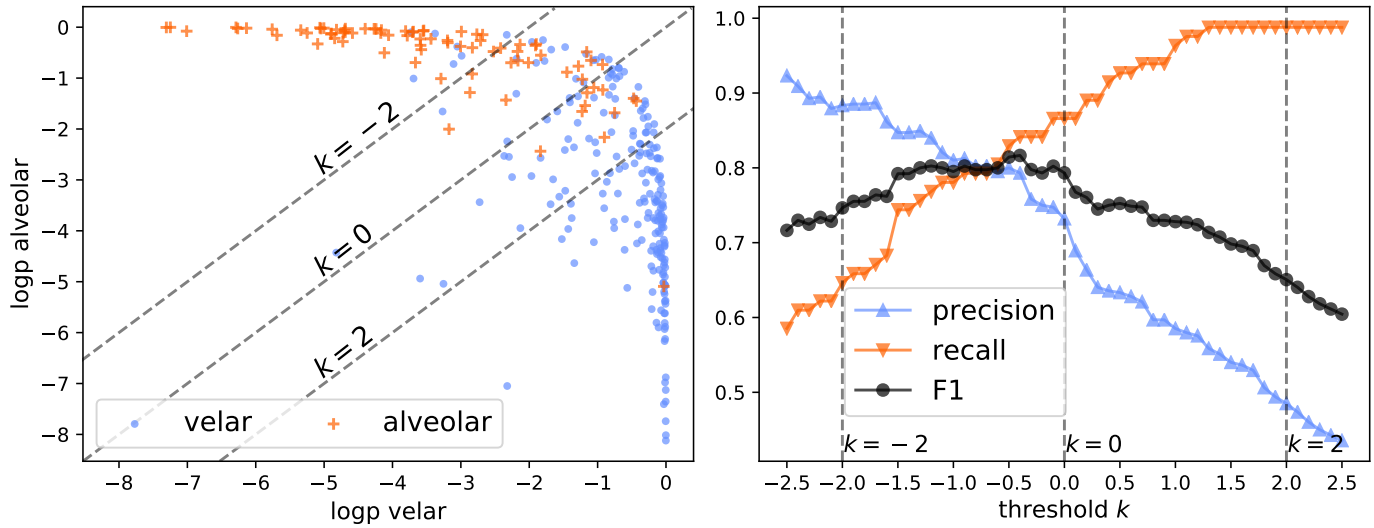

Figure 9: Velar fronting error detection with Audio and Ultrasound system trained on UXTD data. The figure on the left shows model scores for velar and alveolar classes with diagonal lines indicating possible thresholds. Each sample is coloured according to its true label, as given by expert annotators, using $k = 0$. On the right, precision, recall, and F_1 score as a function of the threshold k.

the processes used by annotators to score the samples.

Future work for automatic error detection should explore error processes beyond substitutions, such as insertions or deletions. These cases could be detected using methods similar to those used for mispronunciation detection in language learning (e.g. Witt & Young (2000); Witt (2012); Hu et al. (2015)). There are various child speech corpora which could complement this type of analyses, mostly through the addition of out-of-domain acoustic data. A recent study has identified probe placement variability in the Ultrasuite data (Csapó & Xu, 2020). Variable probe placement could be limiting the performance of an error detection classifier. Future work could leverage the techniques proposed by Csapó & Xu (2020) to account for such variability at training and test time. This could also be used to provide real-time probe placement feedback to clinicians and minimise misalignment errors. Alternatively, to reduce domain mismatch between training and test data, unsupervised domain adversarial training (Ganin et al., 2016) could be helpful, as well as the application of in-domain data augmentation techniques Shorten & Khoshgoftaar (2019). A different direction for future work could could leverage the the temporal dependency of therapy sessions. In this longitudinal online learning scenario (Karanasou et al., 2015), the SLT provides feedback in early sessions (e.g. baseline assessment session) by verifying scores given by the model. Those verified labels are then used to improve model scores on subsequent sessions (e.g. mid-therapy or post-therapy). This scenario could help account for annotator preferences, as well as variability due to speaker characteristics or hardware configuration.

7. Conclusion

We investigated the use of ultrasound tongue imaging for the detection of velar fronting and gliding of /r/ in Scottish English child speakers. For this task, results indicate that experienced speech and language therapists have near perfect agreement when annotating the correctness of velar speech sounds, but agreement on the correctness of rhotic speech sounds is low.

For automatic error detection, out-of-domain adult data improves results on typically developing speech, but it is less useful when evaluating on disordered speech. Results indicate that velar fronting error detection benefits more from ultrasound than audio, but we observe the best performance when using both modalities. In terms of gliding error detection, results are harder to interpret due to low inter-annotator agreement.

Future research should explore techniques to account for speaker, session, and equipment variability with ultrasound equipment, as well as annotation preferences by speech and language therapists. Nonetheless, the overall performance of the classifier is promising, particularly for velar fronting error detection, with good agreement with experienced speech and language therapists. This evidence suggests there is potential for systems to be integrated into ultrasound intervention software for automatically quantifying progress during speech therapy.

8. License and Distribution

The pronunciation scores obtained in Section 4 are publicly available in the Ultrasuite Repository[3] and are distributed under Attribution-NonCommercial 4.0 Generic (CC BY-NC 4.0). A demo of the data preparation and scoring processes with the best performing model in Table 4 is released as part of the UltraSuite code repository[4] under Apache License v.2.

Acknowledgments

We are grateful to the Speech and Language Therapists who kindly agreed to participate in our data collection. This work was supported by the Carnegie Trust for the Universities of Scotland (Research Incentive Grant number 008585) and the EPSRC Healthcare Partnerships grant number EP/P02338X/1 (Ultrax2020 — http://www.ultrax-speech.org).

References

Ahmed, B., Monroe, P., Hair, A., Tan, C. T., Gutierrez-Osuna, R., & Ballard, K. J. (2018). Speech-driven mobile games for speech therapy: User experiences and feasibility. *International Journal of Speech-Language Pathology*, *20*, 644–658.

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

ASHA (2020). Speech sound disorders: Articulation and phonology (practice portal). URL: https://www.asha.org/Practice-Portal/Clinical-Topics/Articulation-and-Phonology American Speech Language Hearing Association. Retrieved September 2020.

Beatty, K. (2013). *Teaching & researching: Computer-assisted language learning*. Routledge.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics*, *19*, 605–617.

Black, M. P., Tepperman, J., & Narayanan, S. S. (2010). Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 1015–1028.

Boyce, S. E. (2015). Articulatory phonetics for residual speech sound disorders: A focus on /r/. In *Seminars in Speech and Language* (p. 257). NIH Public Access volume 36.

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. (under revision). 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*, .

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.

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical linguistics & phonetics*, *29*, 575–597.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*, 37–46.

Csapó, T. G., & Xu, K. (2020). Quantification of transducer misalignment in ultrasound tongue imaging. In *Proc. Interspeech*.

Dudy, S., Bedrick, S., Asgari, M., & Kain, A. (2018). Automatic analysis of pronunciations for children with speech sound disorders. *Computer speech & language*, *50*, 62–84.

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 *Proc. Interspeech*.

Eshky, A., Ribeiro, M. S., Richmond, K., & Renals, S. (2019). Synchronising audio and ultrasound by learning cross-modal embeddings. In *Proc. Interspeech*.

Fabre, D., Hueber, T., Bocquelet, F., & Badin, P. (2015). Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks. In *Proc. Interspeech*.

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.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*, 2096–2030.

Harrison, A. M., Lo, W.-K., Qian, X.-j., & Meng, H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *International Workshop on Speech and Language Technology in Education*.

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154–166.

Johnson, C. J., Beitchman, J. H., & Brownlie, E. (2010). Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology*, .

Karanasou, P., Gales, M. J., Lanchantin, P., Liu, X., Qian, Y., Wang, L., Woodland, P. C., & Zhang, C. (2015). Speaker diarisation and longitudinal linking in multi-genre broadcast data. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 660–666). IEEE.

Katz, W. F., McNeil, M. R., & Garst, D. M. (2010). Treating apraxia of speech (aos) with ema-supplied visual augmented feedback. *Aphasiology*, *24*, 826–837.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA, USA: Sage publications. 2nd Edition.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. URL: https://repository.upenn.edu/asc_papers/43 Retrieved September 2020.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, (pp. 159–174).

Law, J., Boyle, J., Harris, F., Harkness, A., Nye, C. et al. (2000). Preva-

[3] https://www.ultrax-speech.org/ultrasuite

[4] https://github.com/UltraSuite

lence and natural history of primary speech and language delay: findings from a systematic review of the literature. *International Journal of Language and Communication Disorders*, *35*, 165–188.

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2011). The social stratification of tongue shape for postvocalic /r/ in scottish english1. *Journal of Sociolinguistics*, *15*, 256–268.

Lee, A. S.-Y., Law, J., & Gibbon, F. E. (2009). Electropalatography for articulation disorders associated with cleft palate. *Cochrane Database of Systematic Reviews*, .

Lewis, B. A., Avrich, A. A., Freebairn, L. A., Hansen, A. J., Sucheston, L. E., Kuo, I., Taylor, H. G., Iyengar, S. K., & Stein, C. M. (2011). Literacy outcomes of children with early childhood speech sound disorders: Impact of endophenotypes. *Journal of Speech, Language, and Hearing Research*, .

Li, Q., & Russell, M. J. (2001). Why is automatic recognition of children's speech difficult? In *Seventh European Conference on Speech Communication and Technology*.

McCormack, J., Harrison, L. J., McLeod, S., & McAllister, L. (2011). A nationally representative study of the association between communication impairment at 4–5 years and children's life activities at 7–9 years. *Journal of Speech, Language, and Hearing Research*, .

McLeod, S., & Baker, E. (2017). *Children's speech: An evidence-based approach to assessment and intervention*. Boston, MA: Pearson Education.

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, *27*, 1546–1571.

McLeod, S., Davis, E., Rohr, K., McGill, N., Miller, K., Roberts, A., Thornton, S., Ahio, N., & Ivory, N. (2020). Waiting for speech-language pathology services: A randomised controlled trial comparing therapy, advice and device. *International Journal of Speech-Language Pathology*, (pp. 1–15).

Parnandi, A., Karappa, V., Lan, T., Shahin, M., McKechnie, J., Ballard, K., Ahmed, B., & Gutierrez-Osuna, R. (2015). Development of a remote therapy tool for childhood apraxia of speech. *ACM Transactions on Accessible Computing (TACCESS)*, *7*, 1–23.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., & Perdigao, F. (2018). Mispronunciation detection in children's reading of sentences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*, 1207–1219.

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

Ribeiro, M. S., Eshky, A., Richmond, K., & Renals, S. (2019b). Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions. In *Proc. Interspeech*.

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

Richmond, K., Clark, R., & Fitt, S. (2010). On generating Combilex pronunciations via morphological analysis. In *Proc. Interspeech*.

Richmond, K., Clark, R. A., & Fitt, S. (2009). Robust LTS rules with the Combilex speech technology lexicon. In *Proc. Interspeech*.

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 the 18th International Congress of Phonetic Sciences (ICPhS), Glasgow, 10-14 August 2015*. International Phonetic Association.

Sadeghian, R., & Zahorian, S. A. (2015). Towards an automated screening tool for pediatric speech delay. In *Proc. Interspeech*.

Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, *51*, 948–967.

Scobbie, J. M. (2006). (r) as a variable. *Encyclopedia of language & linguistics, Second Edition*, .

Shahin, M. A., Ahmed, B., Ji, J. X., & Ballard, K. J. (2018). Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features. In *Proc. Interspeech*.

Shivakumar, P. G., Potamianos, A., Lee, S., & Narayanan, S. S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In *WOCCI* (pp. 15–19).

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*, 60.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & 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.

Wang, J., Qin, Y., Peng, Z., & Lee, T. (2019). Child speech disorder detection with siamese recurrent network using speech attribute features. In *Proc. Interspeech*.

Ward, L., Stefani, A., Smith, D., Duenser, A., Freyne, J., Dodd, B., & Morgan, A. (2016). Automated screening of speech development issues in children by identifying phonological error patterns. In *Proc. Interspeech*.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. In *Proc. of IS ADEPT (International Symposium on Automatic Detection of Errors in Pronunciation Training)*.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*, 95–108.

Wren, Y., Miller, L. L., Peters, T. J., Emond, A., & Roulstone, S. (2016). Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study. *Journal of Speech, Language, and Hearing Research*, *59*, 647–673.