# Articulatory-to-Acoustic Inversion Mapping using Single Speaker Bilingual Data

*B222755*

*7447*

Master of Science
Speech and Language Processing
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2023

# Abstract

Previous studies on articulatory-to-acoustic inversion (AAI) mapping mainly focus on English data. However, it is known that different languages have different phoneme inventories, and therefore the oral cavity is used differently. We are curious whether a trained AAI mapping learns the universal relationship between acoustics and oral physiology, or that the mapping is actually constrained by the training data. Different model types are also a variable that could influence how well a model is learning. This paper attempts to answer whether multilingual data and model type affect AAI performance.

To investigate multilingual and multispeaker applications for AAI, we start with the comparison of bilingual data of a single speaker. This new dataset of Mandarin and English speech data is composed of synchronized audio recordings, tongue ultrasound videos, and lip camera videos. We then extracted articulator features by DeepLabCut (DLC), a software that tracks the articulatory points with pre-trained models. DLC has not been used before on AAI tasks, and we will discuss its appropriateness.

Two acoustic representations, three variants of long short-term memory (LSTM) models, and three training subsets make up a total of 18 models for us to compare and see which combination has better performance on English and Mandarin test sets. We found that only the language type of training dataset has a significant effect on the root mean square error of model predictions, as bilingual data helps predict both English and Mandarin. Bidirectional LSTM models have the best performance, but the differences among model types are not significant. Our data collection also verifies that DLC is applicable to AAI tasks.

## Acknowledgements

I would like to express my appreciation to all those who made this dissertation and my studies possible.

My deepest gratitude to my family, who supported me financially and mentally over this year. This year has been the most challenging yet most rewarding year in my life. I could not have made it without your support.

I would also like to wholeheartedly thank my advisor, Prof. Korin Richmond, for helping me conceptualize and develop my preliminary ideas, especially for providing the necessary apparatus and guidance during the recording session. His precious advice on "KISS" (*keep it simple stupid*) is one important lesson I'll never forget.

Shout out to Prof. Shan-Shan Wang, Chen-Hsiu Kuo, and Yowyu Lin. I could not have come so far without your lectures in linguistics and your recommendations.

My honor and pleasure of meeting the talented peers in SLP, the most hard-working group of people I've ever encountered. Your endeavors motivate me to keep up with your brilliance.

Thanks to all the staff in the SLP courses and PPLS who helped me along the way and gave me the knowledge to write this dissertation.

And, during this short yet long year, everything that makes me relate to Taiwan, including my friends Shichyo Chou, Jeffrey Ho, and Alex Chen. Also, creators such as LNG Workshop, cbotaku, Ninomae Ina'Nis, and IU accompany my work time. Your presence casts my anxiety and depression away.

# Table of Contents

# Chapter 1

# Introduction

Articulatory-to-acoustic inversion mapping, short for AAI, is a topic widely discussed. AAI estimates the positions of articulators, such as tongue contours or lip shapes, from the acoustic signals. AAI can be applied to many scenarios, such as audio-visual synthesis [Mawass et al., 1997], [Kjellström and Engwall, 2009], computer-aided language learning (CALL) [Badin et al., 2010], and computer-aided pronunciation training (CAPT) [Engwall et al., 2006], [Jones, 2017].

Many model types and acoustic signal representations have been discussed before [Richmond, 2002], [Siriwardena et al., 2022], [Udupa et al., 2023], and [Cho et al., 2023]. However, to the best of our knowledge, there is no paper discussing how different languages and different model types affect the performance of AAI. In this work, we explore the interactions between acoustic representations, languages, and model types. Our choice of model architecture is long short-term memory, whose three variants are used. The training data includes English, Mandarin, and their mixture, while the models are only tested on monolingual data.

We use ultrasound and camera videos to record the movements of articulators, then track points in two-dimensional space using DeepLabCut [Mathis et al., 2018]. The acoustic representations of recordings are the inputs of the models, and the articulator coordinates are the outputs.

With a new dataset and new model types, there are too many uncontrollable variables. Therefore, this work first compares the model performance with previous works [Liu et al., 2015]. We train the model using the mngu0 dataset [Richmond et al., 2011], which is a known AAI dataset with line spectral frequencies acoustic as input and electromagnetic articulography as output training data. A reasonable result on the mngu0 data would suggest that the models are feasible for an AAI task. Next, we use the same model architecture on our newly recorded bilingual dataset. Finally, the performance of models is reported in root mean square errors and Pearson product-moment correlations.

We investigate the effects and interactions of acoustic representations, model types, and training language subsets on model performance. We also discuss the feasibility of using videos and DeepLabCut for AAI tasks, and we build a simple pipeline of raw data processing.

# Chapter 2

# Background

Speech can be seen from two perspectives. One is treating speech as acoustic signals, which are the waveforms transmitted between the speaker and the listener. Another is viewing the articulatory gestures, which are visually available to the listener only if they are looking at the speaker. Though motor theory [Liberman and Mattingly, 1985] suggests that humans recognize speech from the vocal tract gestures, the visual information is still not directly inferrable from the sole acoustic signals.

Articulatory-to-Acoustics Inversion (AAI) mapping aims to provide answers to the question, but one of the most common difficulties AAI encounters is that the relationship between acoustic features and articulatory gestures is one-to-many. For example, the same acoustic signals can be uttered through different coordinations of multiple articulators. This makes the AAI problem speaker-dependent, or even utterance-dependent if the speaker has variations in their own speech.

Another limitation in AAI studies is that the language is mostly restricted to English, as other languages are rare in literature, and only serve as a comparison to English [Wieling et al., 2017]. We wish to look into the effects of training language types, especially languages whose phoneme inventory is quite different from English.

Finally, most studies of AAI use EMA (electromagnetic articulography) data, which is accurate but more infeasible for natural speech data collection. In this paper, we use ultrasound and lip videos to record the articulators, then use DeepLabCut [Mathis et al., 2018] to track the articulator movements. We would like to investigate the appropriateness of this method and establish a workflow for future work with similar demands.

## 2.1 Previous studies on AAI

### 2.1.1 Speaker dependency

The difference between with or without speaker dependence affects a lot. Since everyone has different oral cavities, speaker-dependent models can achieve higher accuracy than speaker-independent ones. Most works use known datasets that are composed of multiple speakers, therefore they train speaker-independent models.

Single speaker studies have the state-of-the-art performance of around 1.881 mm root mean square error and 0.90 correlation [Wang et al., 2022] using the Haskins Production Rate Comparison (HPRC) database [Tiede et al., 2017], while the performance drops to 1.917 mm and 0.89 correlation in multi-speaker for the same database.

A broader scope of speaker dependency is accent dependency, or even language dependency. Wieling and colleagues (2017) compare the interactions among speakers, languages (English and Dutch), and accents. They find a slight reduction in performance across language than within language. This shows that even between two West Germanic languages, AAI has a disadvantage in cross-language prediction. We are curious about what happens when the phoneme inventories are more drastically different, and whether training with two languages helps prediction in both languages.

### 2.1.2 Data collection approaches: EMA and others

The most popular datasets used for AAI include the Multichannel Articulatory (MOCHA) corpus [Wrench, 1999], the HPRC database [Tiede et al., 2017], the Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus [Ji et al., 2014], the mngu0 dataset [Richmond et al., 2011], the Tongue and Lip (TaL) corpus [Ribeiro et al., 2021], etc. The first three datasets use electromagnetic articulography (EMA) measurements of the tongue and lips. Though the accuracy of EMA is higher than ultrasound since the sensors are directly connected to the articulators, its invasiveness nature causes the subjects to speak unnaturally.

The TaL corpus uses ultrasound and lip videos to collect the data points of tongue position and lip shape. This approach is non-invasive and cheaper than EMA, which is desired for natural, inexpensive data collection. Nonetheless, its downsides include the possibility of shifting the apparatus across sessions and the extra process of tracking points from video files. Therefore, this approach might be more prone to errors.

Previous research has used DeepLabCut [Mathis et al., 2018] to extract feature points from videos and transformed the coordinates as training data for silent speech recognition [Beeson and Richmond, 2023]. However, no previous studies have adopted this computer vision approach to AAI tasks. We would like to test the feasibility of DeepLabCut so that future data collection will be much easier through ultrasound and camera videos.

### 2.1.3 Acoustic feature types: MFCC, LSF, and SSL representations

Mel-frequency cepstral coefficients, or MFCC, are representations of speech data by collectively making up a representation of the power spectrum of a waveform using Fourier Transform and Mel scale transformation. Then, the power is taken log and treated as a signal ("spectrum of a spectrum"). Each set of MFCCs can be seen as a slice of the spectrogram. More detailed MFCCs have the velocities ("deltas") and accelerations ("delta-deltas") of the feature vector, which provide additional temporal information about the trajectories of features.

Another representation is line spectral frequencies (LSF). The mngu0 dataset [Richmond et al., 2011] has 40 LSF and the gain of 10 surrounding frames of the current EMA frame. LSF is favored due to its insensitivity to noise, but MFCC is proven to perform better on AAI tasks [Ghosh and Narayanan, 2010].

Convolutional neural networks (CNN) can also represent speech information due to their ability to encode information about surrounding frames and other features. Illa and Ghosh (2019) found that an end-to-end network composed of a CNN for acoustic feature extractor and a BLSTM for AAI mapping model had a better RMSE performance on the MOCHA-TIMIT corpus.

With the recent prominence of self-supervised learning (SSL) models, acoustic signals are found to be better represented. SSL models have different objectives, input features, and predictions, but they can be broadly dichotomized into two methods: predictive and contrastive. Predictive SSL aims to minimize the differences between representations, while contrastive SSL focuses on distinguishing samples from different objects but fails at encoding complete information. Common SSL models include PASE+ [Ravanelli et al., 2020], wav2vec [Schneider et al., 2019], TERA [Liu et al., 2021], AALBERT [Chi et al., 2021], Mockingjay [Liu et al., 2020], DeCoAR [Ling et al., 2020], etc. In a Transformer-based AAI task, TERA is found to perform best in most model size conditions [Udupa et al., 2023].

### 2.1.4 AAI model types: MLP, RNN, and Transformer

Early research on AAI has used multilayer perceptrons (MLP) and mixture density networks (MDN) to approach this problem [Richmond, 2002], which has shown that MLP is possible to perform AAI while MDN can better solve the non-uniqueness in AAI.

Recurrent neural networks (RNN) such as long short-term memory (LSTM) and gated recurrent units (GRU) have often been implemented on speech data due to their ability to store long dependency information, which is applicable to sequence data like speech. A GRU has two gates to reset and update information, while an LSTM unit has three gates, namely input, output, and forget gates. Therefore, GRUs need fewer training parameter, cost less memory, and executes faster than LSTMs, while LSTM is more accurate on a larger dataset. Previous studies of AAI RNN models are comparable on the MOCHA-TIMIT dataset [Liu et al., 2015], [Wu et al., 2023] and the HPRC dataset [Shahrebabaki et al., 2019], [Siriwardena et al., 2022].

Transformer architectures are also favorable due to their powerful processing abilities. [Udupa et al., 2023], [Cho et al., 2023]. However, the demands on computing resources and huge amounts of training data make Transformers applicable to certain scenarios, such as AAI datasets and computation capacities that are large enough.

The directionality of models is also a hyperparameter. If the model is bi-directional, it takes the entire sequence as input and predicts the output of a certain frame. More context means higher accuracy, but bi-directional models are not capable of generating predictions from ongoing speech.

### 2.1.5 Evaluation metrics: RMSE and PPMC

Common metrics on how well an articulatory-to-acoustic inversion mapping model performs include root mean square error (RMSE) between the predicted and ground truth values. The measurement is often de-normalized into real-life units, such as pixels or millimeters. Another evaluation is Pearson product-moment correlation (PPMC, or coefficient correlations *CC*), which measures the linear relationship between two variables.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(e_i - \bar{e})^2}{n}} \tag{2.1}$$

$$r = \frac{\sum_i^n (e_i - \bar{e})(t_i - \bar{t})}{\sqrt{\sum_i^n (e_i - \bar{e})^2 \sum_i^n (t_i - \bar{t})^2}} \tag{2.2}$$

$e_i$ is the model prediction and $t_i$ is the ground-truth features. $\bar{e}$ is the average of predicted values and $\bar{t}$ is the average of ground-truth values.

Subjective evaluation includes human preferences for animated outputs. The ground truth and predicted values are plotted to points in images, or even built as a 3D module of articulators, then animated as videos. However, this evaluation is too time-consuming for a dissertation, which is why we only consider objective evaluations, i.e., RMSE and PPMC.

## 2.2 This paper

There are only a few papers discussing how different languages affect the performance on AAI [Wieling et al., 2017], not to mention that, to the best of our knowledge, no previous papers compare two very different languages, i.e., English and Mandarin, and their effects on the performance of different model types [1].

English and Mandarin have distinct characteristics. One of the major differences is that Mandarin has tones, which affect the supralaryngeal articulations to some degree. Other than tones, Mandarin has voiceless and voiced alveolo-palatal fricatives /ɕ,ʑ/, voiceless and voiced retroflex fricatives /ʂ,ʐ/, voiced labio-palatal approximant /ɥ/, and phonemic contrast between consonant aspiration instead of voicing in English.

### 2.2.1 Three variants of LSTM models

For input acoustic feature types, we choose Mel Frequency Cepstral Coefficients (MFCCs) as the input representation of each frame. 13 MFCCs are extracted to represent the acoustic signal corresponding to a video frame. We also include MFCCs with their velocities and accelerations (deltas and delta-deltas) as another feature type, which is 39-dimension.

As for the model's output, we use ultrasound and camera videos to record the articulator movements, and use DeepLabCut [Mathis et al., 2018] to capture the coordinates of points on the articulators.

[1] Note that the EMA-MAE corpus has Mandarin-accented speakers speaking English, but no Mandarin data.

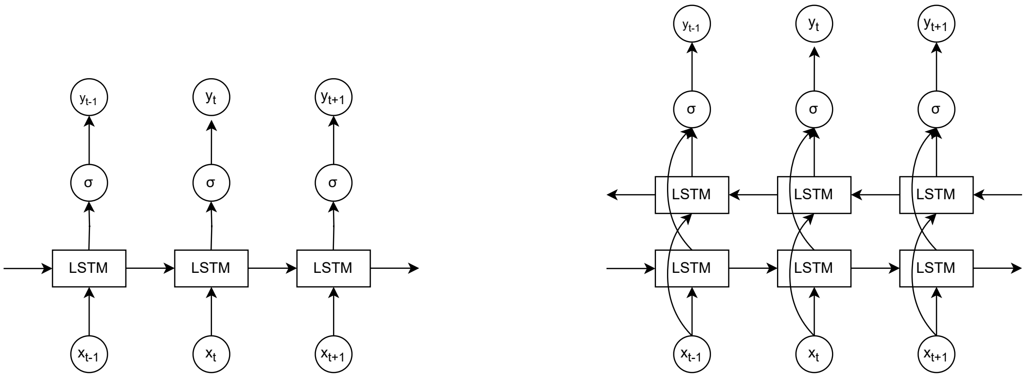
We use three variants of LSTM models, which are long short-term memory (LSTM) models, Context Window (CW) LSTMs, and bidirectional LSTM (BiLSTM) models.

Long short-term memory models are composed of LSTM units, each consisting of a cell, an input gate, an output gate, and a forget gate. The inputs of a cell are the current observation and previous LSTM cell information. The three gates learn how much information to take from the inputs and how much to output to the next unit, meaning that each cell can remember previous values of arbitrary time.

Our second model, CW, uses a context window with LSTM models, making a longer input dimension. CW models can be useful in pragmatic applications; imagine a situation of making a 3D figure move its lips in response to input speech. We want to make the model accurately predict the ultrasound features in real-time, thus a practical model cannot be bi-directional, for it needs the entire speech sequence to output the prediction of the current frame. Our LSTM model can do the trick due to its uni-directionality; nevertheless, we want to push the performance as much as we can. It is known that audiovisual asynchronicity within 100 ms can be calibrated by human brains [Vroomen et al., 2004]. That is why we are using 6 future frames (= 100ms in 60fps) to help predict the output features of the current frame. Take note that the context window here is defined as only right-ward; it does not include past frames since LSTM should already encode previous information.

BiLSTM models are two LSTM models, one right-ward and one left-ward, stacked together. This enables the model to look at the whole input sequence and therefore should have the best performance.

The LSTM and BiLSTM model architecture is illustrated in Figure 2.1.



(a) LSTM model

(b) BiLSTM model

Figure 2.1: Core model architecture of LSTM variants. The $\sigma$ represents the sigmoid function, while each "LSTM" is a long short-term memory unit.

The model performance is evaluated by RMSE and PPMC.

### 2.2.2 Research questions: effects of models and languages

We have two main research questions:

1. Which model type is the best for AAI?
2. What are the effects of adding multiple languages into training?

Other issues arise along with the progression of this paper, which we will also discuss:

1. How efficacious is ultrasound useful for AAI?
2. Is DeepLabCut point tracking reliable for AAI?

# Chapter 3

# Method/Experimental Setup

## 3.1 Recording

### 3.1.1 Subject

The subject has agreed to our participant information sheet and data management plans, and the experiment procedures have been approved by the School of Informatics at the University of Edinburgh.

The speaker subject is not a professional, but his proficiency in English and Mandarin is desired. He is a native speaker of Mandarin and has near-native proficiency in English. To be more specific, his English accent is American, and his Mandarin accent is Taiwanese.

We hope this dataset can be expanded in the number of subjects, languages, and accents in the future.

### 3.1.2 Reading text selection

#### 3.1.2.1 English

For the English reading text, we use the prompts from the existing TaL corpus [Ribeiro et al., 2021], which contains sentences from various sources, including accent identification paragraph [Weinberger, 2015], the Rainbow Passage [Fairbanks, 1940], the Harvard sentences [Rothauser, 1969], the TIMIT corpus [Garofolo, 1993], the VCTK corpus [Yamagishi et al., 2019], and the Librispeech corpus [Panayotov et al., 2015].

To see how much of the English phoneme inventory is represented in the TaL prompt, we use Festival [Taylor et al., 1998] to transcribe this English prompt and check its diphone coverage. For the whole reading prompt, there are 90 unique uniphones and the unique diphone count is 1287, which constitutes 15.89% of the total coverage. However, due to constraints of time and effort, we did not finish recording the entire prompt, and the uniphone coverage we achieved is 83, and the diphone count is 1113, which is 13.74% of all possible diphones.

#### 3.1.2.2 Mandarin

Mandarin has 21 onsets and 36 rimes, which include monophthongs, diphthongs, and nasal codas. Due to phonotactic constraints, there are 408 combinations of onsets and rimes, and 1317 combinations if different tones are counted separately. The list of pinyin combinations takes reference from the Chinese Xinhua Dictionary (`http://xh.5156edu.com/pinyi.html`, accessed on 17 August 2023), where pinyins with and without tones are specified. Then, we manually filter out words that serve as an auxiliary purpose, such as exclamation or question particles. Finally, some pinyin combinations are dialectal and therefore do not exist in the Taiwanese Mandarin dictionary (`https://dict.revised.moe.edu.tw/search.jsp?md=1`, accessed on 17 August 2023). We also remove these words.

Previous research on the tone effects of articulatory gestures in pronouncing Mandarin has shown contradictory results. Supralaryngeal articulations are significantly different when pronouncing monosyllabic words containing bilabial onsets with /a/ (/ba, pa, ma/) in tone 1 or tone 3 [Erickson et al., 2004]. Contrastive evidence indicates that articulator positions remained the same when pronouncing vowels without onsets in 4 different lexical tones [Torng, 2000].

Even though the two studies focus on different materials, such as the existence of onsets and vowel inventory size, they have shown conflicting results. Since the acoustic signal can encode pitch information, we decide to play it safe and include all possible combinations of pinyins and tones in the reading prompt.

After listing out all 1317 combinations of pinyin, we select sentences from the Chinese Gigaword corpus [Parker et al., 2009], which is a corpus of Mandarin newspapers. Specifically, we pick the year 2007 of the cna_cmn category as our text corpus since the subject speaker is Taiwanese, therefore a Taiwanese newspaper will be more familiar to him. A text selection algorithm is designed to pick the maximum counts of new pinyins with the shortest sentence possible. Existing pinyins will not be considered in the next iteration of selection. The words are transcribed by the Pinyin package [Yu, 2016] on Python.

With the above-mentioned algorithm, 439 news texts are included in the Mandarin reading prompt. Some pinyins do not appear in the newspaper, possibly because the genre of journalism limits the vocabulary. We look up the rest of the pinyins in the Revised Mandarin Chinese Dictionary (`https://dict.revised.moe.edu.tw/search.jsp?la=1`, accessed on 17 August 2023) to find a bi-word for each of those pinyins. Then, each bi-word is put inside a carrier sentence *wo shuo* __ 'I said __' for the completion of all possible pinyin combinations in the reading prompt.

The pinyin coverage of the entire prompt is 100%, but due to time limitations, we did not finish recording all the sentences. The current count of pinyin types is 1015, which is 77% of coverage.

### 3.1.3 Apparatus

We use AAA (Articulate Assistant Advanced) [Wrench, 2017] to record three types of data, including audio files (.wav), ultrasound video of tongue surface (.avi), and lip positions (.avi). The subject sits in a hemi-anechoic chamber and reads prompts on a screen in the front. The length of the recording session is 30 minutes so that the speaker does not experience excessive physical and psychological burdens. He can also pause to rest whenever he wishes.

The recording microphone is Sennheiser HKH 800 p48, recording with a 48kHz sampling frequency at 16 bits. The distance between the speaker and the microphone is not fixed, but the speaker is asked to be as close as possible for better recording quality.

The ultrasound probe is Articulate Instruments' Micro system with a 92° field of view. In B-mode ultrasound, there are 64 ultrasound splines, and each spline has 864 pixels. Each spline scans $\sim$ 80fps times per second, and information on all splines contributes to the white contour of the tongue on a fan-shape space, as indicated in Figure 3.1. The probe is aligned to the center of the subject's jaw and is adjusted to a suitable angle in order to capture a midsagittal view of the tongue with as much tongue information as possible.

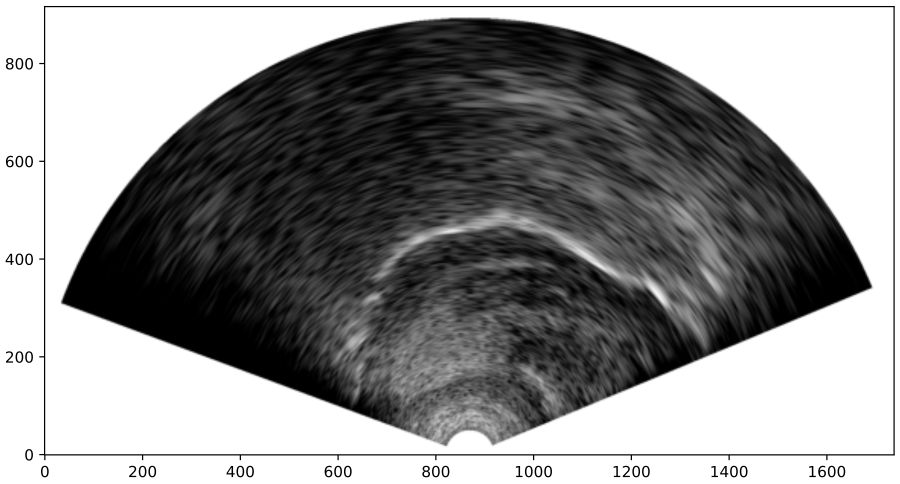
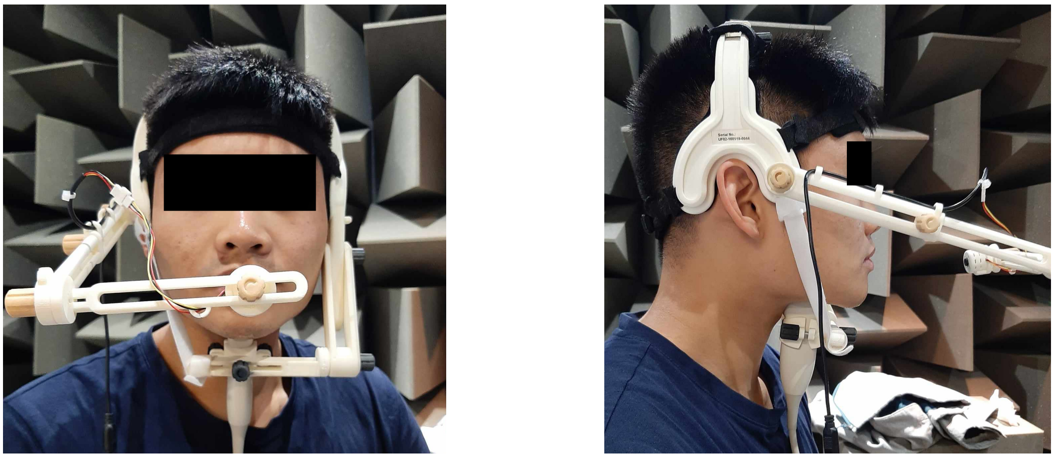


Figure 3.1: Example of an ultrasound slice in the AAA graphic user interface and its exported file.

The camera recording the lips has a frame rate of $\sim$ 60fps recording in greyscale. The camera is extended from the helmet where the ultrasound probe is also fixated on. The relative positions of the speaker, ultrasound probe, and lip camera are illustrated in Figure 3.2.



(a) Front view

(b) Side view

Figure 3.2: Front view and side view of the subject wearing the experiment apparatus.

The three types of data are synchronized through the Articulate Instruments' SynchBrightUp unit [Articulate Instruments Ltd, 2010]. The SyncBrightUp unit flashes a white square in the upper-left corner of the lip videos, and simultaneously sends a beep signal to synchronize with the recording and the ultrasound video. The apparatus setup mostly follows the steps in `https://materials.articulateinstrume` nts.com/Manuals/Installation%20manual%20for%20Micro%20Ultrasound%20system%20Rev%202.pdf (accessed on 17 August 2023). Due to lighting in the hemi-anechoic chamber, some lip videos are too bright that the white flash cannot be detected by AAA, so we re-calibrate the synchronization manually.

### 3.1.4 Post-processing

After exporting audio files from AAA, we get the MFCCs from the waveforms using librosa [McFee et al., 2023], which are 13 MFCCs extracted from 40 Mel frequencies, using a 20ms Hamming analysis window with a 10ms frame shift. We also include MFCCs with their velocities and accelerations (deltas and delta-deltas) as another feature type, which is 39-dimension.

The ultrasound and lip recordings are exported as separate AVI video files with a pixel size of 320*240. Both the ultrasound and lip videos are deinterlaced using Moviepy (https://github.com/Zulko/moviepy, accessed on 17 August 2023) to avoid frame-by-frame stripes. Then, the videos are resampled to 60 frames per second. Finally, extra silences in the start and end of videos are removed, as to shrink data size and to reduce unhelpful noises.

After unifying the formats of the videos, we use DeepLabCut (DLC) [Mathis et al., 2018] to trace the articulators. The vanilla DLC takes a video input and requires human annotators to label the points in a partition of files for DLC to trace other files using machine learning. Still, without labeled data, we can use pre-trained DLC models [Wrench and Balch-Tomes, 2022] to track 14 points of tongue ultrasound and 8 points of lip videos. For each frame's each point, three values are returned, i.e., x coordinates, y coordinates, and the model's confidence in predicting that point. The tongue contour is fitted quite accurately, as Figure 3.3 indicates.

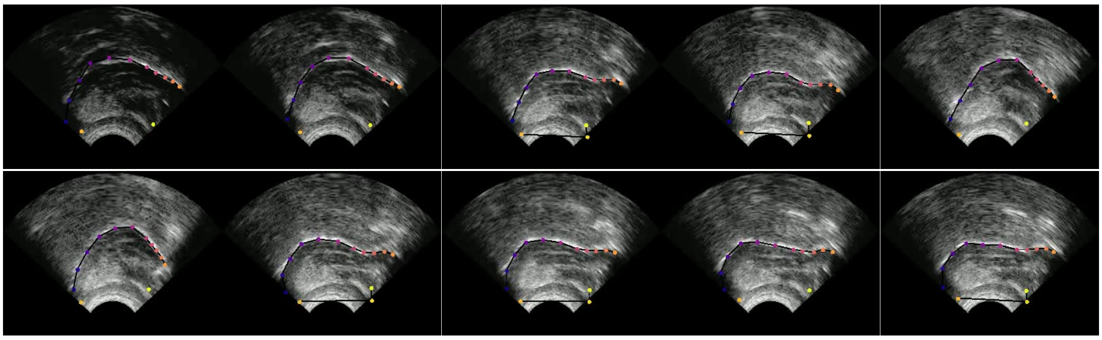


Figure 3.3: A strip of images converted from the DLC video output of tracing the tongue ultrasound. The points are closely tracing the tongue surface. Each picture is a snapshot intervened by 400ms, the order is from the top right corner and left-ward to the bottom right corner. This utterance is “This was easy for us.”

However, due to the brightness and contrast in the lip video, the confidence of point tracking is quite low for some frames (Figure 3.4). Notice the 3rd snapshot where the upper lip is not traced and the 6th snapshot where the lower lip points are incorrectly tracing the upper edge of the lower teeth instead of the lip. We tried manually labeling 20 recordings from each language training subset to fine-tune the pre-trained model, but the results are still as bad.

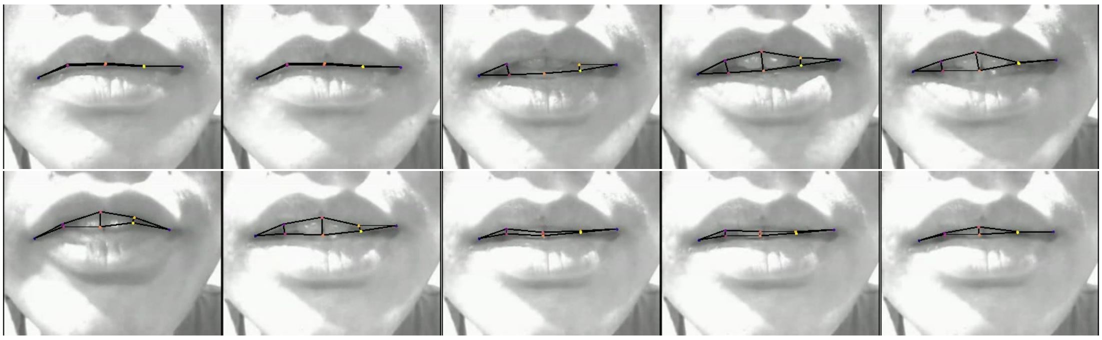


Figure 3.4: A strip of images converted from the DLC video output of tracing the lip video. The points are closely tracing the tongue surface. Each picture is a frame snapshot intervened by 400ms, the order is from the top right corner and left-ward to the bottom right corner. This utterance is also “This was easy for us.”

For points whose confidence values are lower than 0.8, we find the closest point before and after whose confidence values are greater than 0.8. Then, we interpolate the x coordinate value and y coordinate value respectively, by equal intervals between the two closest points. Finally, the interpolated x and y coordinate values replace those uncertain points. The effects of interpolation are illustrated in Figure 3.5, where the first subplot is the x-axis of the top left inner lip, and the second subplot is its y-axis. The values are raw data points, which have not been cleaned by lowpass filtering or normalization.

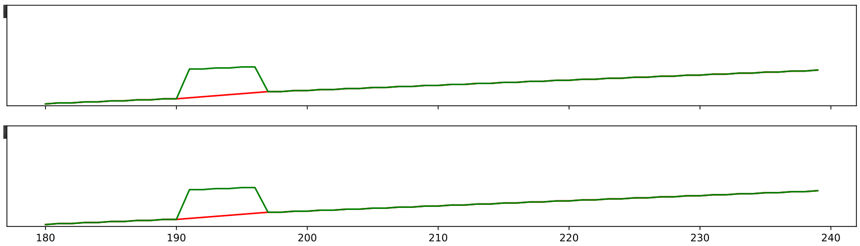


Figure 3.5: Before (green) and after (red) interpolation of the x-axis (above) and y-axis (below) of the top left inner point of the lips. This segment is adapted from the 180th to the 240th frame of the utterance “Jane may earn more money by working hard.”

### 3.1.5 Data cleaning

Richmond (2002) has shown that lowpass filtering of the raw points can re-calibrate the shifts during recording sessions and can moderately improve the performance on AAI. For each feature, we find the mean of each recording and plot the mean values chronologically. We then apply a Butterworth filter with an order of 3 and a critical frequency of 0.05 using SciPy [Virtanen et al., 2020]. The lowpass filter is shown in Figure 3.6.

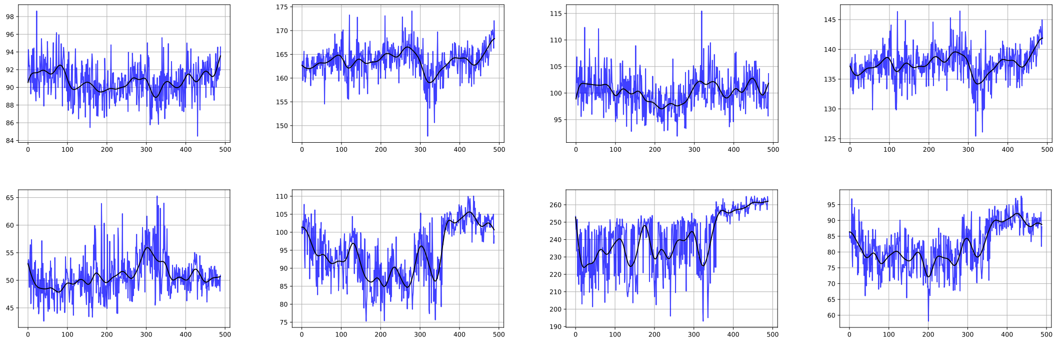


Figure 3.6: Lowpass filters on 4 ultrasound features (*vallecula_x, vallecula_y, tongueRoot1_x, and tongueRoot1_y*) and 4 lip features (*leftLip_x, leftLip_y, rightLip_x, and rightLip_y*). The dataset is English recordings.

We can see that there is a dip around the 320th recording in all articulators, which is probably caused by a shift of apparatus between sessions.

Then, we normalize the data points by subtracting the filter value from each frame value. In other words, for a specific recording, its frames are compared with respect to its filter value. This simple method makes all recordings more comparable without shifts. Finally, we z-score normalize each feature by subtracting the respective global mean and then dividing by the respective global standard deviation.

For each frame of every recording, the input vectors are 13-dimension for MFCC features, or 39-dimension if we're using MFCC with deltas. The corresponding 28 ultrasound features and 16 lip features are concatenated as 44-dimension output vectors. Since each recording has different lengths, the number of frames is different. For ease of comparison across subsets, we zero-pad each array to the maximum length of all recordings, which is 2667 frames. In other words, all MFCC input vectors are (2667,13), and MFCC with deltas are (2667,39). Afterward, we z-score normalize each of the 13 features respectively, using the global mean and standard deviation across audio files and timesteps. This is to avoid features (especially the first feature "energy", the red line in Figure 3.7) with larger oscillations over time from dominating other features.

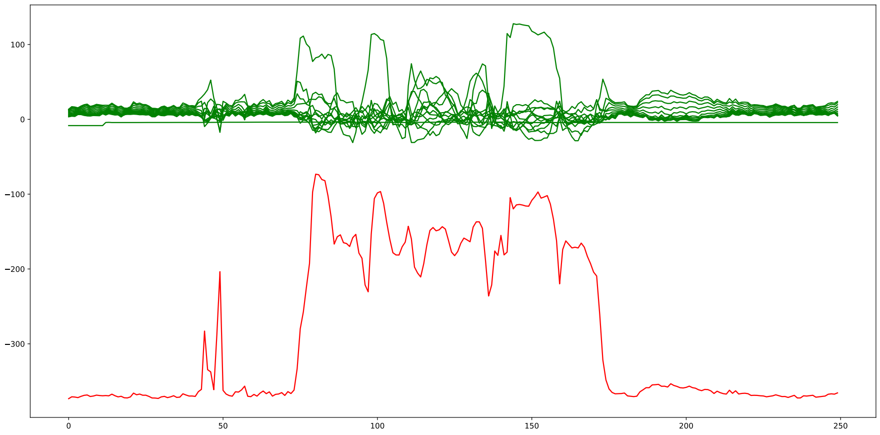


Figure 3.7: 13 MFCCs across timesteps of one recording, which is the first in the English dataset, reading "This was easy for us." The red line is the energy, which ranges from -400 to -100. In comparison, other features range from around -50 to 100.

Initially, 300 training data and 10 testing data are randomly selected from 488 English and 381 Mandarin utterances. For the bilingual train set, we combine each of the first 150 data from the English and Mandarin dataset. Across different acoustic feature representations, the training and testing recordings are kept the same.

## 3.2 Model training

### 3.2.1 Comparison with a known dataset *mngu0*

Since this dataset is newly created, we cannot guarantee how well the models perform on it. To ensure that the models are not faulty by themselves, we use the mngu0 dataset [Richmond et al., 2011] to test their performance.

In comparison with Liu and colleagues' DBLSTM model [Liu et al., 2015], our feedforward layers have a unit size of 128 instead of 300, and only one BiLSTM layer with 64 units instead of two layers with 100 units. This lowers the number of parameters, so a worse performance is expected but should not differ a lot.

We use root mean square error (RMSE, Equation 2.1) and Pearson product-moment correlation (PPMC, Equation 2.2) as evaluation metrics. Each EMA feature has its metrics calculated separately. The RMSE and correlation values of the test results are reported in Table 3.1.

Table 3.1: RMSEs and PPMCs of the test set, predicted from our BiLSTM and LSTM model. The unit of RMSE is in millimeters (mm).

| body parts | RMSE | | PPMC | |
|---|---|---|---|---|
| | BiLSTM | LSTM | BiLSTM | LSTM |
| T3_x | 1.227 | 1.119 | 0.839 | 0.872 |
| T3_y | 1.595 | 1.435 | 0.888 | 0.917 |
| T2_x | 1.372 | 1.231 | 0.842 | 0.879 |
| T2_y | 1.346 | 1.144 | 0.919 | 0.945 |
| T1_x | 1.424 | 1.196 | 0.849 | 0.895 |
| T1_y | 1.506 | 1.236 | 0.903 | 0.937 |
| JAW_x | 0.665 | 0.562 | 0.797 | 0.862 |
| JAW_y | 0.816 | 0.718 | 0.903 | 0.930 |
| UL_x | 0.368 | 0.314 | 0.760 | 0.835 |
| UL_y | 0.505 | 0.448 | 0.782 | 0.819 |
| LL_x | 0.718 | 0.605 | 0.838 | 0.897 |
| LL_y | 1.189 | 1.053 | 0.885 | 0.915 |

T1-3 are tongue tip, body, and dorsum, respectively. UL is the upper lip, and LL is the lower lip. The suffixes x/y are the coordinates of the body part.

Since Liu and colleagues (2015) summarize the RMSE and PPMC values across testing data and channel features, we also take the same approach and compare their models' results against our models'.

In their paper, DNN (Deep Neural Network) is their baseline model, which is composed of three feedforward layers of 300 units, while the input has a context window of 10 frames, corresponding to 5 frames prior to and 5 frames after the predicted EMA frame. DBLSTM (Deep Bidirectional Long Short-Term Memory) uses only the 7th frame of the input context window, which is 15ms after the corresponding EMA frame. The input features are passed to two feedforward layers of 300 units, which serve as a feature extractor. Two bidirectional LSTM layers followed as the output layer. DRMDN (Deep Recurrent Mixture Density Network) replaces the DBLSTM model's output layer with one mixture density output layer.

We have a similar BiLSTM model as Liu and coworkers', but we use the 6th frame, which is 10ms after the predicted EMA frame. Our model also used two feedforward networks but with a smaller size of 128. Furthermore, we have only one bidirectional LSTM in our BiLSTM model and two LSTM layers in our LSTM model. This is to keep the number of parameters comparable, as our discussion of different models' performance would like to consider only their directionality, so the model sizes are better kept similar. We use 80% of the mngu0 data for training and 10% each for validation and test set.

The first three models are Liu and peers', and our models are the last two.

Table 3.2: RMSEs and PPMCs of the mngu0 test set in Liu et al., (2015) and this paper. The unit of RMSE is in millimeters (mm).

| author | models | RMSE (mm) | PPMC |
|---|---|---|---|
| Liu et al., 2015 | DNN | 1.000 | 0.869 |
| | DBLSTM | 0.816 | 0.921 |
| | DRMDN | 0.832 | 0.914 |
| this paper | BiLSTM | 0.922 | 0.892 |
| | LSTM | 1.061 | 0.851 |

By testing the model architecture on a known dataset mngu0, we verify that the models are working as expected. We now use the models to train on our dataset.

### 3.2.2 Model architecture

We use Keras [Chollet et al., 2015] to build our deep learning models. The sequential models are trained on NVIDIA V100 (Volta) GPU cards. All models take the input of (2667,$n$), where $n$ is different according to different features and architectures.

The first layer is a masking layer to remove zero-paddings in individual recordings. Then two feedforward layers followed to serve as a feature extractor. The intermediate layers are three types of LSTM layers. The output layer is a dense layer to output 44 articulatory features. The shared model architecture is illustrated in Figure 3.8.

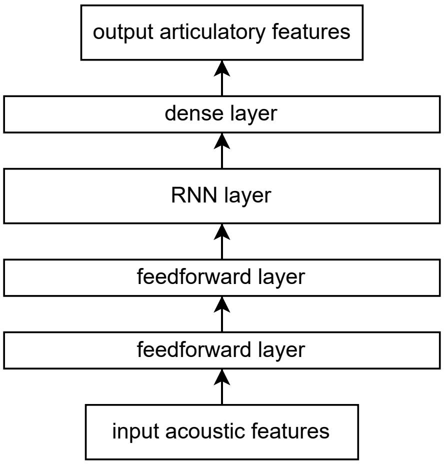


Figure 3.8: Overview of the model structure shared by three model types. The input dimension is $(13n, timsteps)$, where $n$ depends on whether the model is a CW model and whether the MFCC input has deltas. The output dimension is $(44, timesteps)$.

The training objective is by mean square error (MSE), and the optimizer is Adam [Kingma and Ba, 2017]. The batch size is 32 and is trained for 100 epochs.

### 3.2.3 Parameter size

For our BiLSTM models, we have one BiLSTM layer with 128 units. This is to make the number of parameters comparable with the other two model types. The CW models take future 6 frames as input, so the input MFCC feature vector for each frame is $13 \times 6 = 78$, or $39 \times 6 = 234$ if we're using MFCC with deltas. For LSTM models, there are two unidirectional LSTM layers, each with 128 units as well. Each model's number of parameters is recorded in Table 3.3.

Table 3.3: Number of parameters by acoustic feature and model types.

| MFCC types | BiLSTM | CW | LSTM |
|---|---|---|---|
| with deltas | 126,124 | 131,884 | 106,924 |
| without deltas | 122,796 | 111,916 | 103,596 |

# Chapter 4

# Results

The root mean square error (RMSE) is calculated by averaging the RMSE between each predicted and ground-truth tensor. First, we remove zero-paddings from the ground-truth tensors, and the predicted tensors are sliced to corresponding sizes. For one testing data, the array size is (44, None), where None is the total number of frames in 10 testing data. For each 44 body part features, the RMSE is calculated independently by Equation 2.1 between prediction and ground-truth vectors. After iterating through the features, we take the average of all RMSE values as the result for a model evaluated on a test set.

Note that the RMSE unit is not in millimeters because there is no statement scale in the lip recordings, so the measurement is in pixels (px) from the original video resolution 320*240.

For each of the 44 articulatory features, we use Pearson product-moment Correlation (Equation 2.2) to see the correlation between the predicted and ground-truth curves. Each feature calculates a PPMC value by taking 10 recordings of a test set as one long time series. The resulting PPMC averages across 44 features.

The RMSE and PPMC results are presented in Table 4.1 and Table 4.2.

Table 4.1: RMSEs and PPMCs of three model types, trained on three train sets, and tested on two languages. The input vectors are MFCC *without* their deltas and delta-deltas. The bold-faced value is the best performance across three model types.

| | | BiLSTM | | CW | | LSTM | |
|---|---|---|---|---|---|---|---|
| train set | test set | RMSE | PPMC | RMSE | PPMC | RMSE | PPMC |
| Eng | Eng | **2.487** | 0.994 | 2.569 | 0.994 | 2.516 | 0.994 |
| Eng | Mand | **4.789** | 0.992 | 5.024 | 0.996 | 5.035 | 0.994 |
| Mand | Eng | 8.232 | 0.984 | 7.019 | 0.986 | **6.890** | 0.985 |
| Mand | Mand | 1.198 | 0.998 | **1.188** | 0.998 | 1.254 | 0.997 |
| Both | Eng | 2.858 | 0.993 | 3.005 | 0.992 | **2.835** | 0.993 |
| Both | Mand | **1.266** | 0.997 | 1.335 | 0.997 | 1.320 | 0.997 |

Table 4.2: RMSEs and PPMCs of three model types, trained on three train sets, and tested on two languages. The input vectors are MFCC *with* their deltas and delta-deltas. The bold-faced value is the best performance across three model types.

| | | BiLSTM | | CW | | LSTM | |
|---|---|---|---|---|---|---|---|
| train set | test set | RMSE | PPMC | RMSE | PPMC | RMSE | PPMC |
| Eng | Eng | 2.803 | 0.993 | 2.743 | 0.994 | **2.571** | 0.994 |
| Eng | Mand | **4.887** | 0.994 | 5.043 | 0.994 | 5.177 | 0.994 |
| Mand | Eng | 7.847 | 0.985 | **7.096** | 0.986 | 7.390 | 0.986 |
| Mand | Mand | 1.259 | 0.997 | 1.244 | 0.997 | **1.194** | 0.998 |
| Both | Eng | **2.795** | 0.993 | 2.824 | 0.993 | 2.834 | 0.993 |
| Both | Mand | **1.256** | 0.997 | 1.323 | 0.997 | 1.316 | 0.997 |

The evaluation metrics, RMSE and PPMC, have shown good performance of the models. High PPMCs demonstrate that the LSTM models map the articulatory features from acoustic signals, as the models' predictions fit with the ground-truth curves. The gaps between ground-truth and prediction are evaluated by RMSE, so we only look further into the RMSEs in search of meaningful patterns among acoustic representations, model types, and training language subsets.

Looking deeper into the RMSE values across model types, we notice that the BiLSTM model is not always the best out of the three models in every training and testing dataset scenario. This shows that model types do not affect much on the model performance, as further ANOVA test also indicates.

As for influences of training languages, models trained on one language perform the best when tested on the same language subset. Still, the bilingual model achieves high performance in both test sets without substantially increasing errors. We conclude that AAI is language-dependent, and training with bilingual data helps boost performance when tested in both languages.

## 4.1 ANOVA test on RMSE

To test whether there are significant differences among different model types and languages, we perform a three-way ANOVA test. We collect the RMSEs of 44 output features in 2 test subsets, English and Mandarin. We then label each row with three factors, "feature" (two levels, "MFCC" and "MFCC_deltas"), "model" (three levels, "BiLSTM", "LSTM", and "CW"), and "language" (three levels, "English", "Mandarin", and "bilingual"). Therefore, the table consists of $44 \times 2 \times 2 \times 3 \times 3 = 1584$ rows. The ANOVA table is illustrated in Table 4.3.

Table 4.3: ANOVA results of different acoustic features, model types, and training language subsets. The asterisks (*) in the table indicate significance. "sum_sq" = "sum of square", "df" = "degree of freedom", and "F-value" represents the variation between sample means.

| | sum_sq | df | F-value | p-value |
|---:|---:|---:|---:|---:|
| C(feature) | 1.4e-20 | 1.0 | 1.7e-21 | 1.0e+00 |
| C(model) | 7.5e+00 | 2.0 | 4.7e-01 | 6.3e-01 |
| C(language) | 1.4e+03 | 2.0 | 8.6e+01 | *2.7e-36 |
| C(feature):C(model) | 1.6e-23 | 2.0 | 1.0e-24 | 1.0e+00 |
| C(feature):C(language) | 1.2e-21 | 2.0 | 7.7e-23 | 1.0e+00 |
| C(model):C(language) | 4.3e+01 | 4.0 | 1.3e+00 | 2.5e-01 |
| C(feature):C(model):C(language) | 3.0e-22 | 4.0 | 9.4e-24 | 1.0e+00 |
| Residual | 1.2e+04 | 1566.0 | NaN | NaN |

There are no significant interactions between factors, and only "language" has a significant effect on RMSE values. The unimportance of acoustic representations fits our intuition, as AAI can often achieve good performance by using the simplest 13 MFCCs. With higher dimensions of deltas and delta-deltas, it might increase undesired noises in input data. Model types perform similarly well is unexpected, as we hypothesize that bidirectional models should learn better than the other two. Nevertheless, it might be that all three models learn equally well, and using the simplest architecture "LSTM" model will be sufficient in this AAI task.

To look deeper into the significant factor "language", we pool the non-significant factors "feature" and "model" and re-run the model. The result is shown in Table 4.4.

Table 4.4: ANOVA results after removing insignificant interactions. The asterisks (*) in the table indicate significance. "sum_sq" = "sum of square", "df" = "degree of freedom", and "F-value" represents the variation between sample means.

| | sum_sq | df | F-value | p-value |
|---|---|---|---|---|
| C(language) | 1371.690 | 2.0 | 86.804 | *1.7e-36 |
| Residual | 12491.617 | 1581.0 | NaN | NaN |

The p-value of "language" is less than 0.05, meaning that training language subsets have statistically significant effects on RMSE values. To see which language affects the model performance, we run a post-hoc Tukey's honest significance difference (HSD) test on the RMSE values grouped by language, as demonstrated in Table 4.5.

Table 4.5: Tukey's HSD test of different training language subsets.

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|---|---|---|---|---|
| English - Both | 9.800 | 0.000 | 3.823 | 15.777 |
| Mandarin - Both | 13.162 | 0.000 | 7.185 | 19.139 |

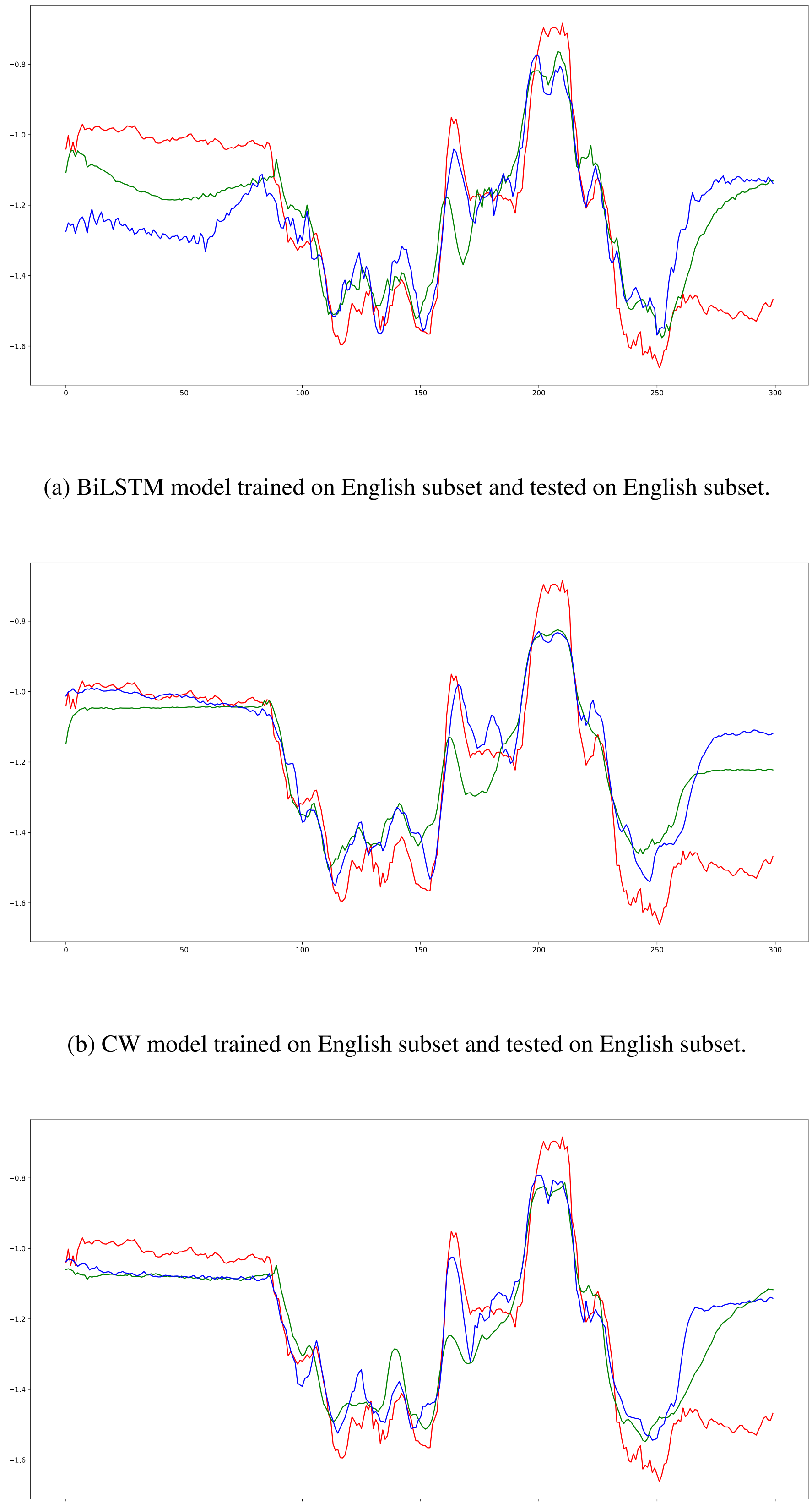
Table 4.5 shows that the difference between English and Mandarin training subsets is not significant. The bilingual training subset has the lowest overall RMSE, which is reasonable as models trained on bilingual data learn more about the diverse phoneme inventory of both languages. We conclude that AAI is language-dependent, and training with bilingual data helps boost performance when tested in both languages.

Yet a looming shadow is that the bilingual model is actually learning the two global means from the English and Mandarin train set, which makes the overall testing RMSE lower since the model predicts between the two means. To rule out this possibility, we require more data and a finer normalization procedure in future work.

In summary, the acoustic representation and model types are insignificant as the grouped post-hoc test in Table 4.3 indicates, and only the language factor is significant between each language and bilingual data.

## 4.2 Visualization of ground-truth values vs. model predictions

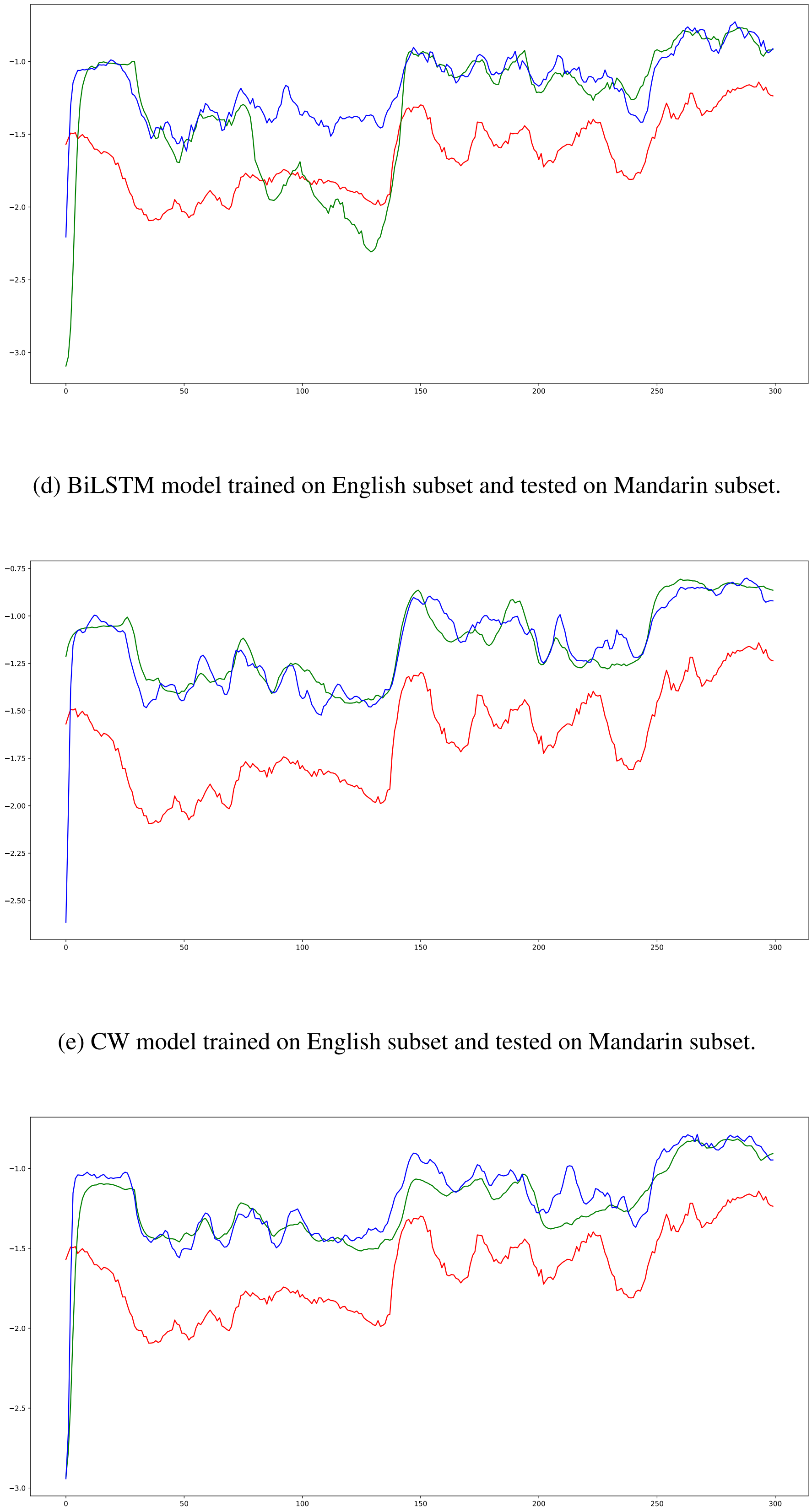
To see how the models predict each channel, we adopt a more fine-grained analysis of the 6th output feature, which is the x coordinate of the tongue body. This feature is chosen because it reflects more movements of articulators than stable parts of the tongue, such as the tongue root or vallecula. Figure 4.6 presents how each model predicts the x coordinate of tongue body in English and Mandarin test sets.



(a) BiLSTM model trained on English subset and tested on English subset.

(b) CW model trained on English subset and tested on English subset.

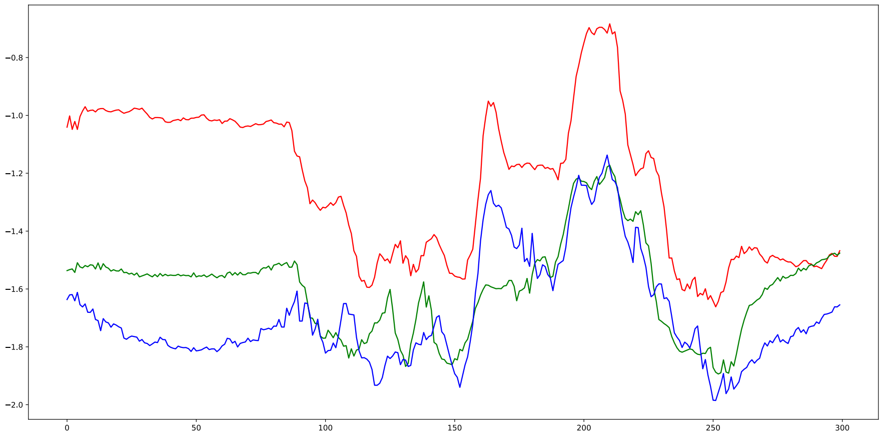
(c) LSTM model trained on English subset and tested on English subset.



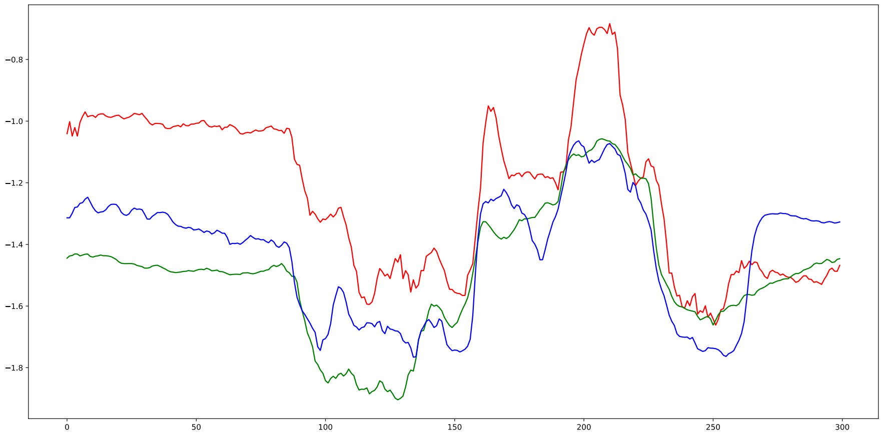
(d) BiLSTM model trained on English subset and tested on Mandarin subset.

(e) CW model trained on English subset and tested on Mandarin subset.

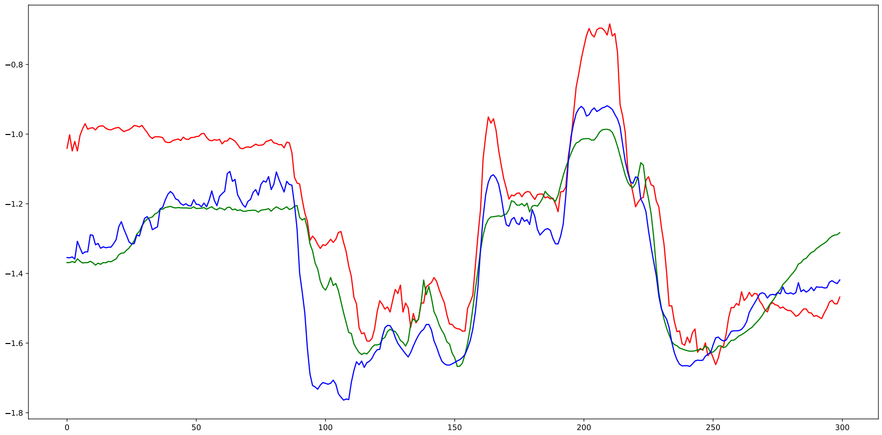
(f) Models trained on English subset and tested on Mandarin subset.



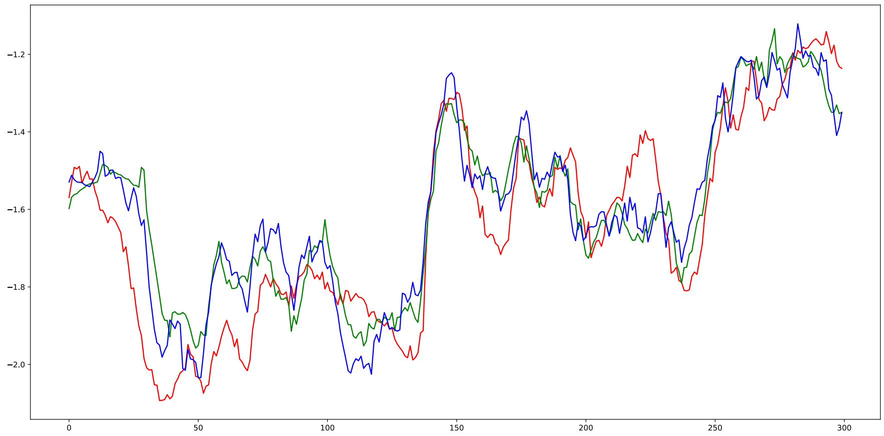
(g) BiLSTM model trained on Mandarin subset and tested on English subset.



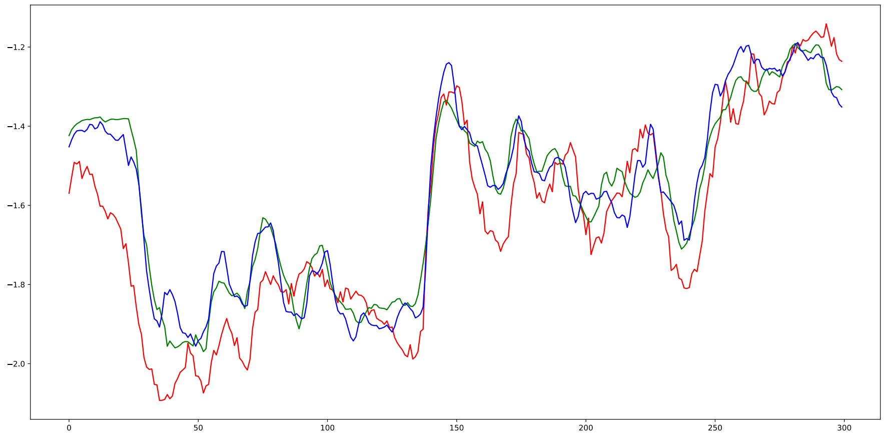
(h) CW model trained on Mandarin subset and tested on English subset.



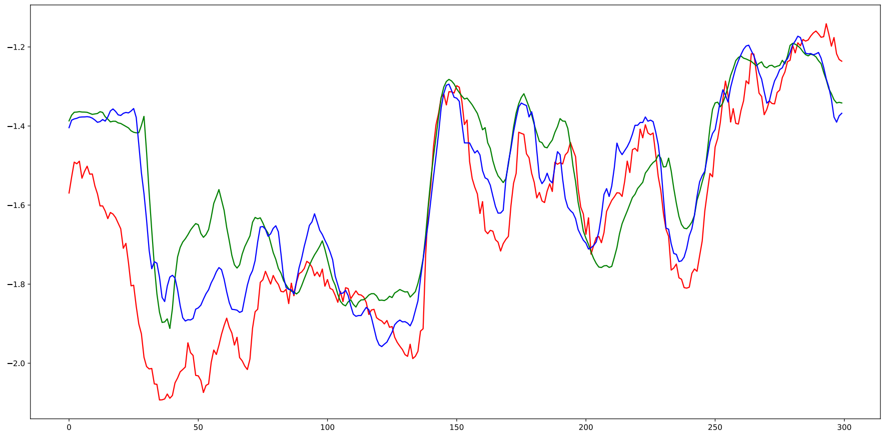
(i) LSTM model trained on Mandarin subset and tested on English subset.



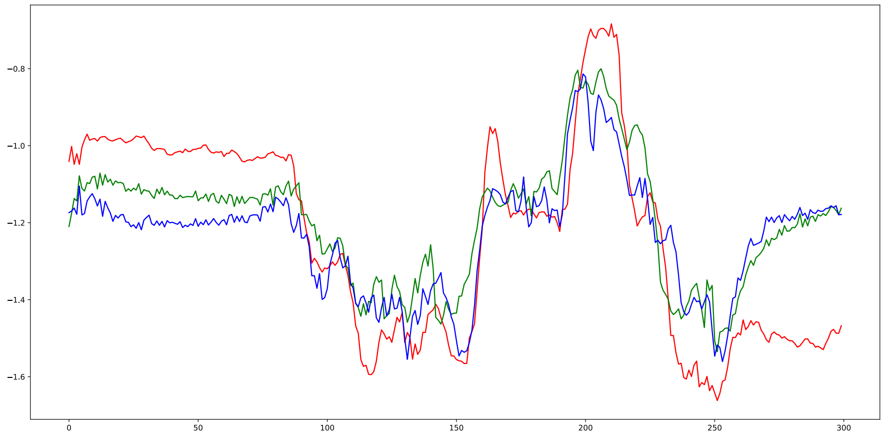
(j) BiLSTM model trained on Mandarin subset and tested on Mandarin subset.



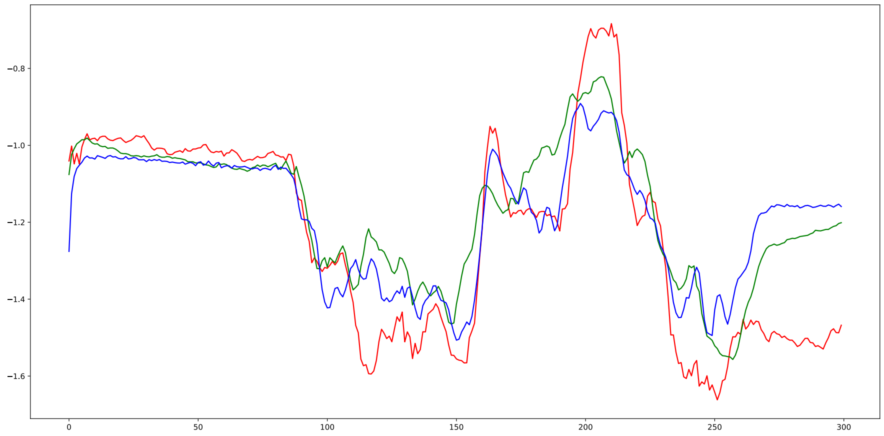
(k) CW model trained on Mandarin subset and tested on Mandarin subset.



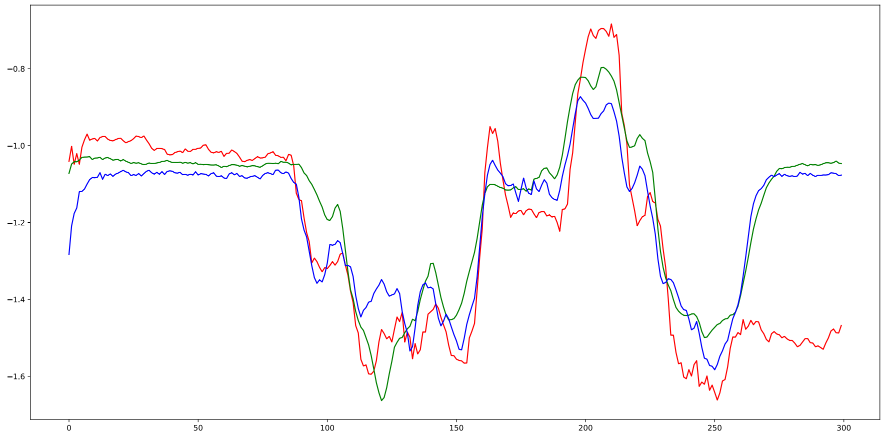
(l) LSTM model trained on Mandarin subset and tested on Mandarin subset.



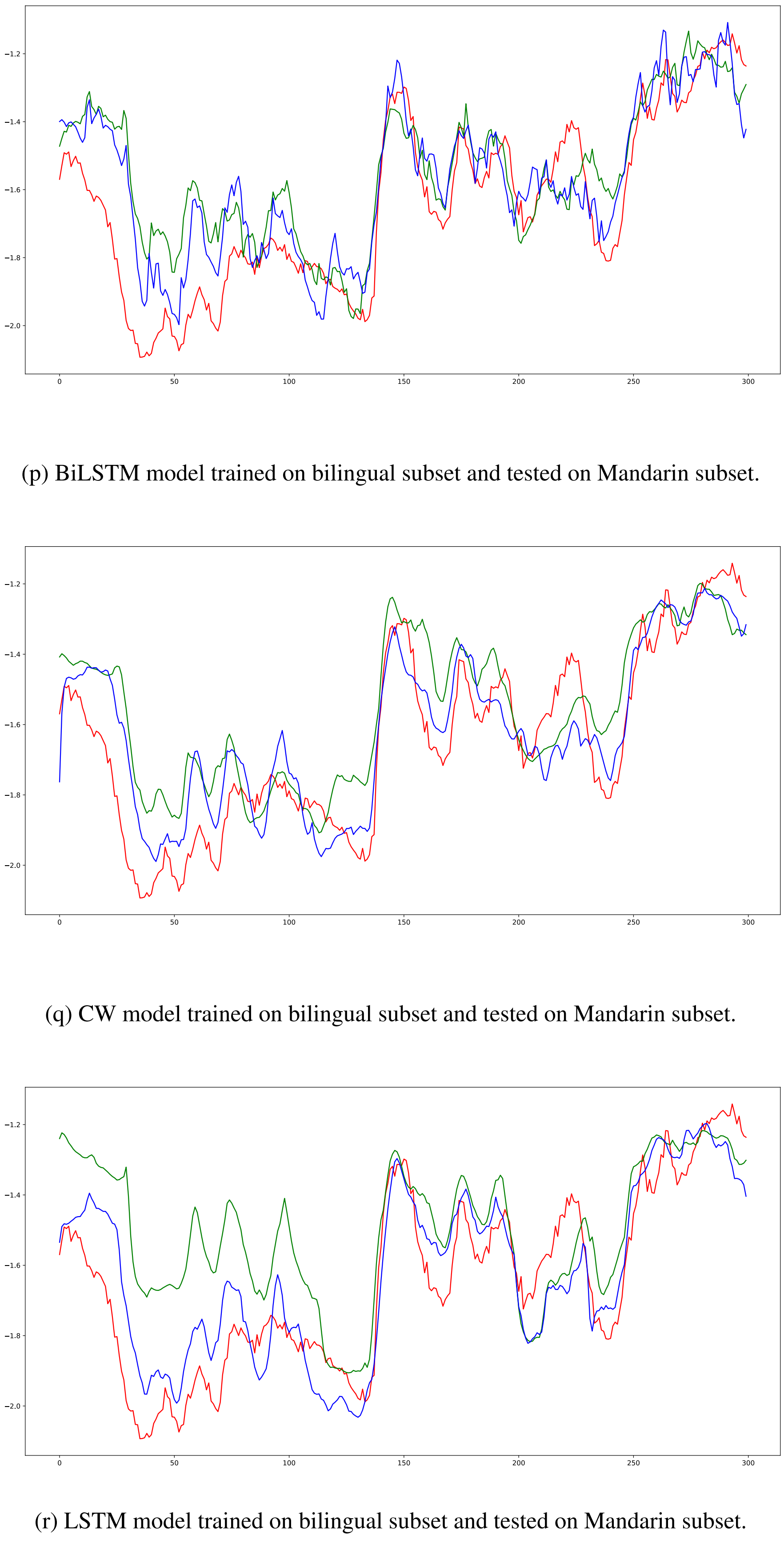
(m) BiLSTM model trained on bilingual subset and tested on English subset.



(n) CW model trained on bilingual subset and tested on English subset.



(o) LSTM model trained on bilingual subset and tested on English subset.



(p) BiLSTM model trained on bilingual subset and tested on Mandarin subset.

(q) CW model trained on bilingual subset and tested on Mandarin subset.

(r) LSTM model trained on bilingual subset and tested on Mandarin subset.

Figure 4.6: Curves of ground-truth and predictions from different models. The red curve is the ground-truth, the predictions of models trained on MFCC features without deltas are plotted in green, and those trained on MFCC deltas are in blue. The English testing utterance is the first 5 seconds of the 14th recording, and the Mandarin testing utterance is the first 5 seconds of the 20th recording.

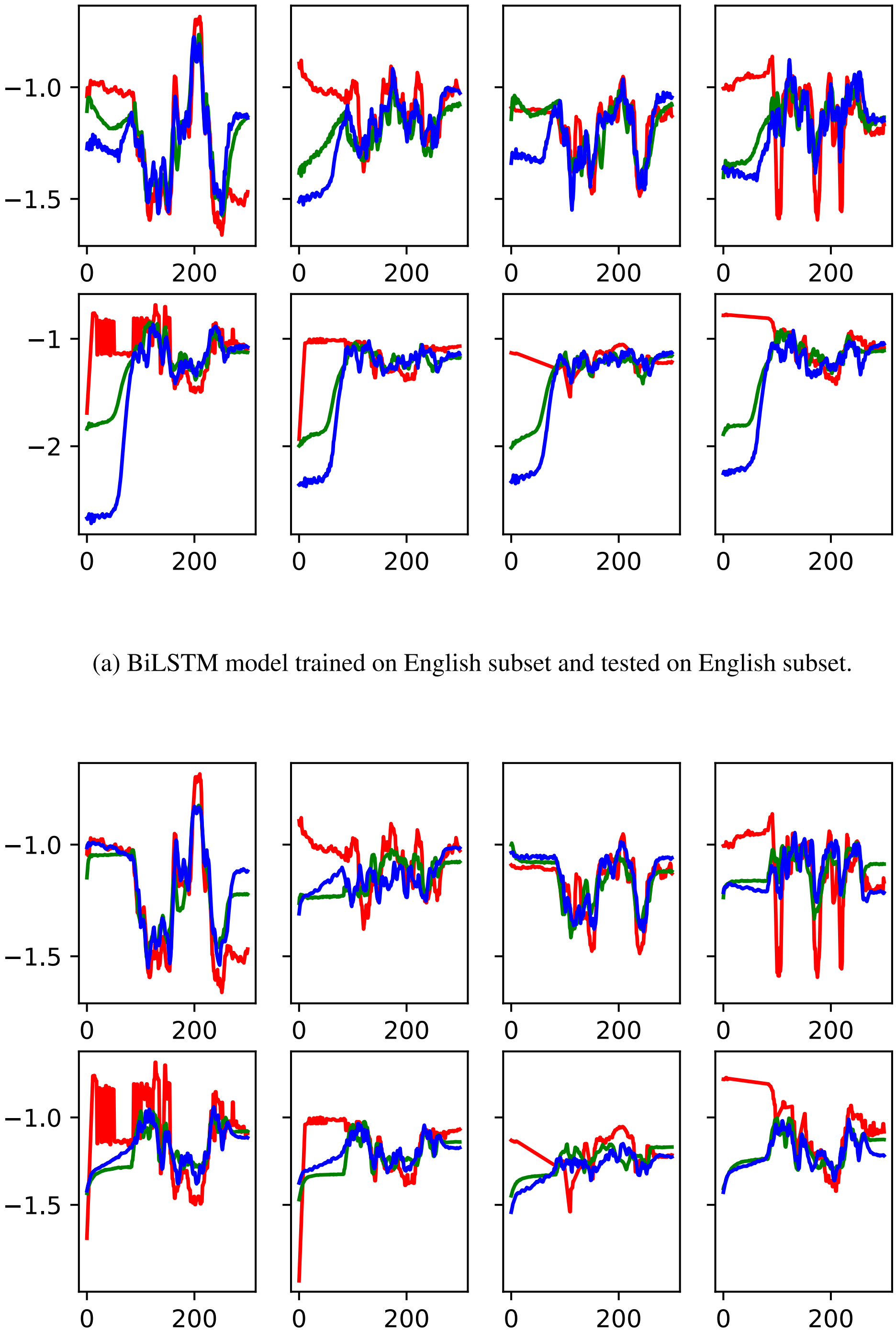
We can see that the predicted curves fit pretty well to the ground-truth values, while there are some complex relationships between acoustic representations and model types.

The green and blue curves represent predictions trained from MFCCs acoustic representations, with and without deltas. There is not a recognizable pattern between the two curves, as Table 4.3 also suggests insignificant differences between the two acoustic representations.

Subfigures (d)-(i) have clear offsets between the ground-truth values and predictions. This is because the train set and test set are different languages. To be more specific, when the train set is English and the test set is Mandarin (subfigures (d)-(f)), the prediction is larger; on the other hand, when the train set is Mandarin and the test set is English (subfigures (g)-(i)), the prediction is smaller. We suggest that this offset might be language-inherent, as the mean of articulator features in Mandarin is found to be constantly smaller. We will further discuss this in Chapter 5.1.1.

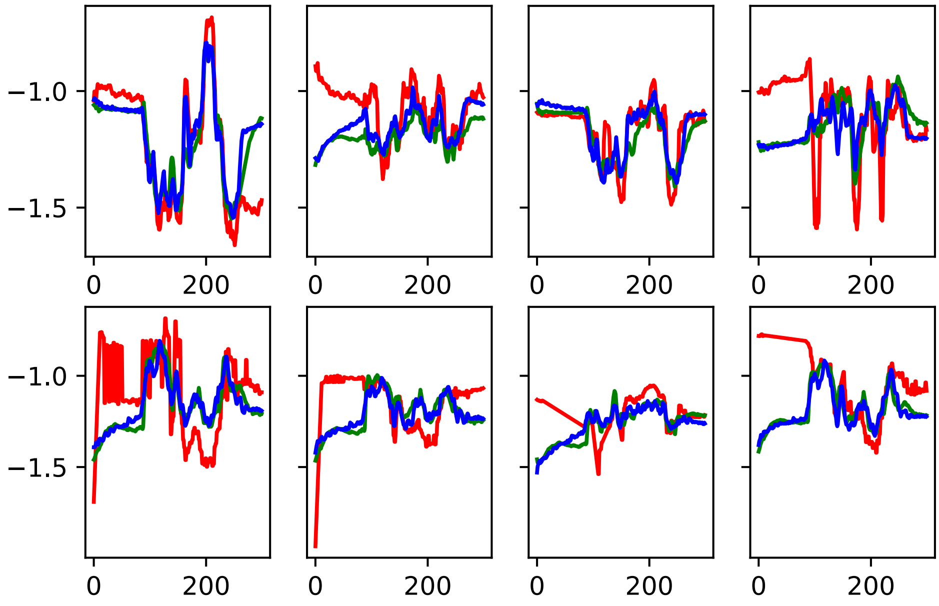
Notice a peculiar plateau appears at the start of many model predictions, which corresponds to a short silence at the beginning of the recording. The articulators are moving toward the intended positions as the ground-truth curves illustrate, but there is no acoustic signal for the model to make accurate predictions. Under this circumstance, the model is only trained to predict a fixed value to minimize the error.

After seeing how the models fit one feature, an overview of more features presents the ground-truth and prediction features, 4 from tongue ultrasound (x and y coordinates of the tongue body and tip) and 4 from lip videos (x and y coordinates of the left and right corner of the lip) in Figure 4.15. Note that the phrase “ground-truth” in this paper hereafter is defined as the coordinate results from DLC tracking, whose reliability is not fully verified, and therefore is not an actual *ground truth* such as the EMA data mentioned above.

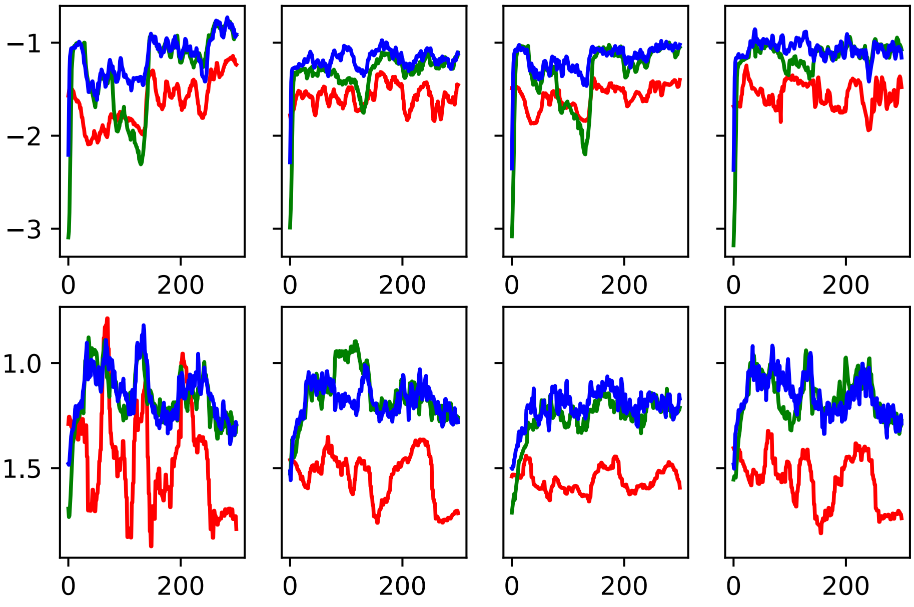


(a) BiLSTM model trained on English subset and tested on English subset.

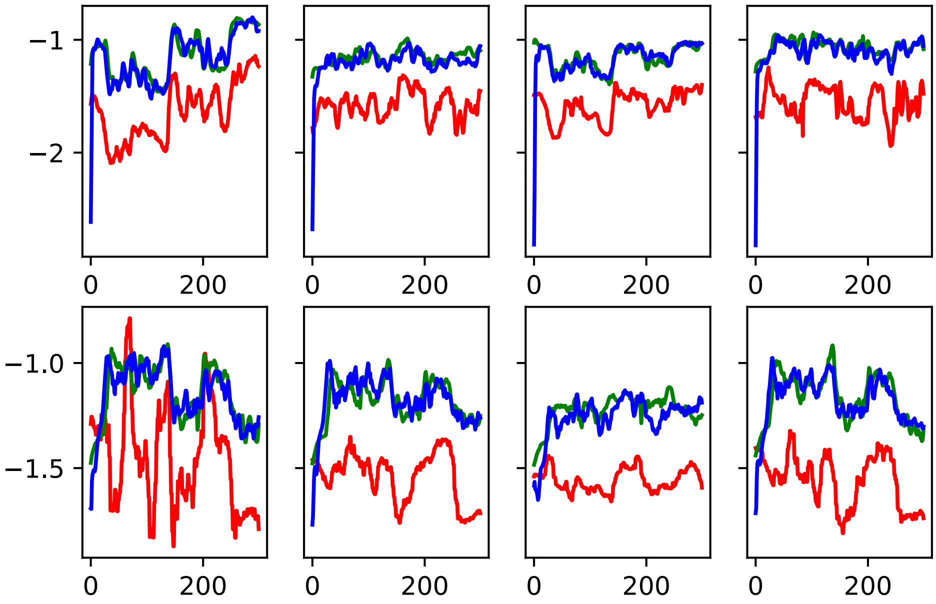
(b) CW model trained on English subset and tested on English subset.



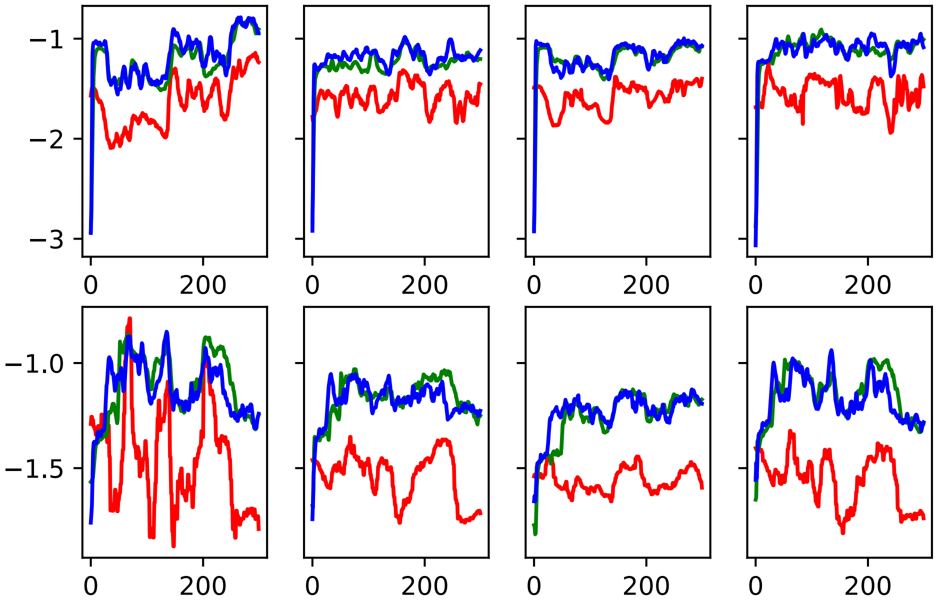
(c) LSTM model trained on English subset and tested on English subset.



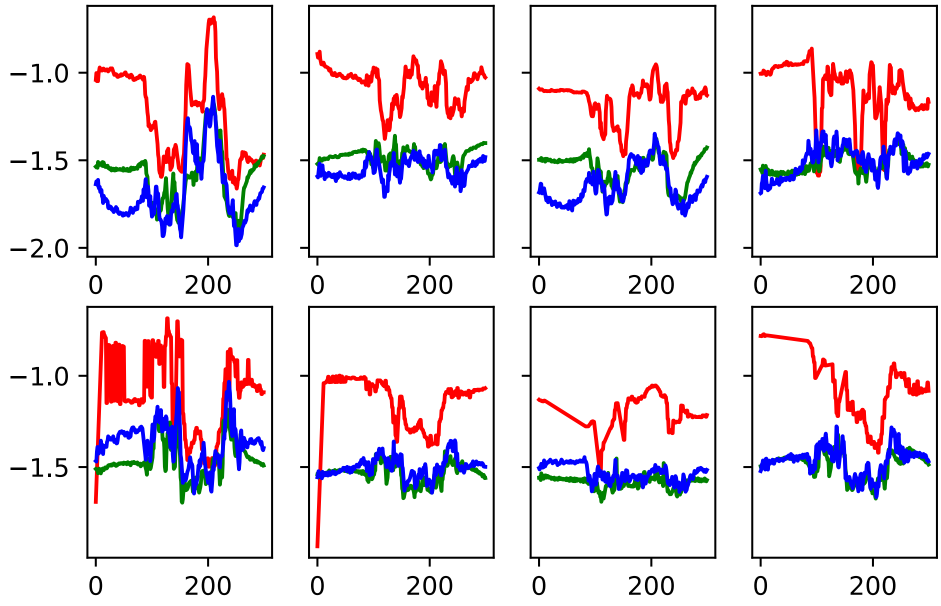
(d) BiLSTM model trained on English subset and tested on Mandarin subset.



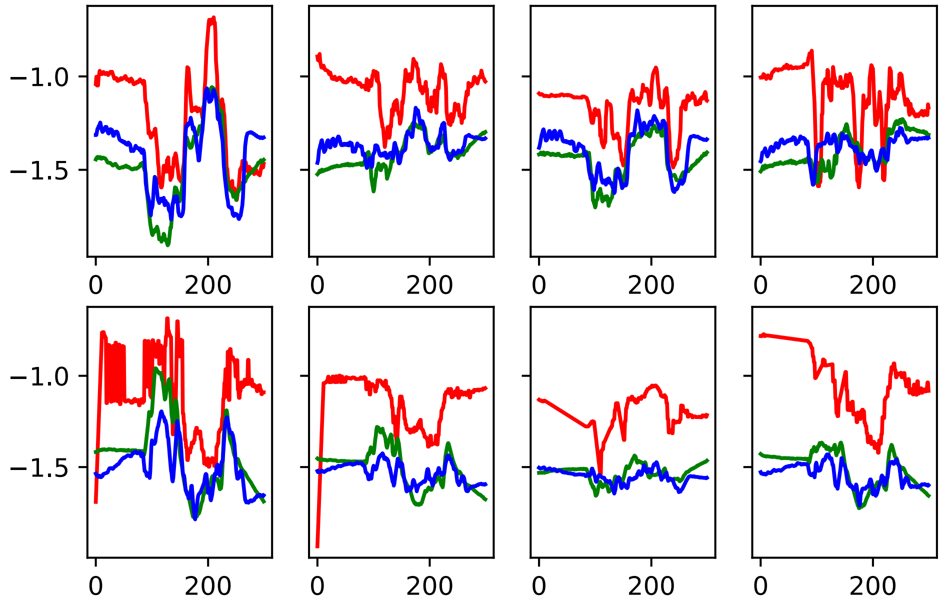
(e) CW model trained on English subset and tested on Mandarin subset.



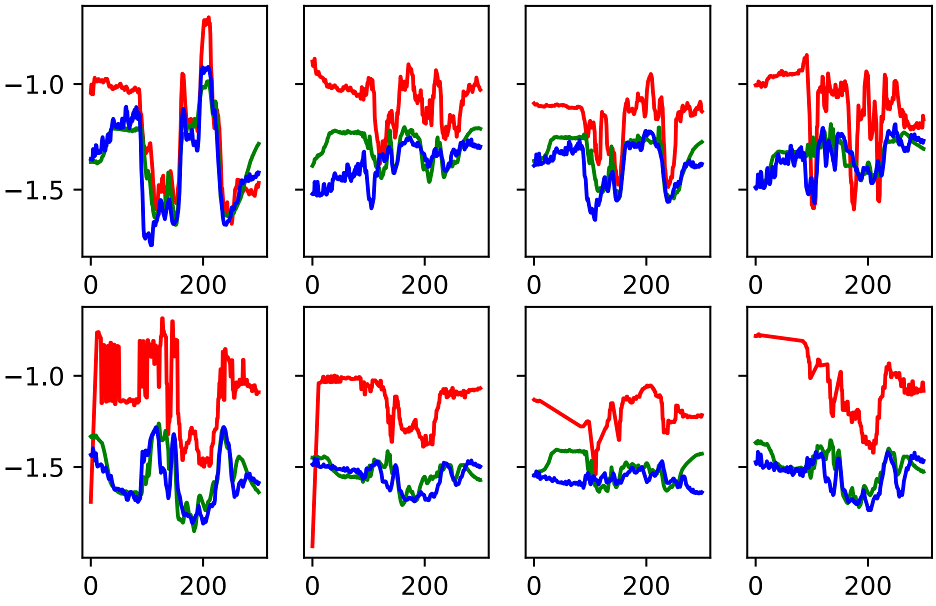
(f) LSTM model trained on English subset and tested on Mandarin subset.



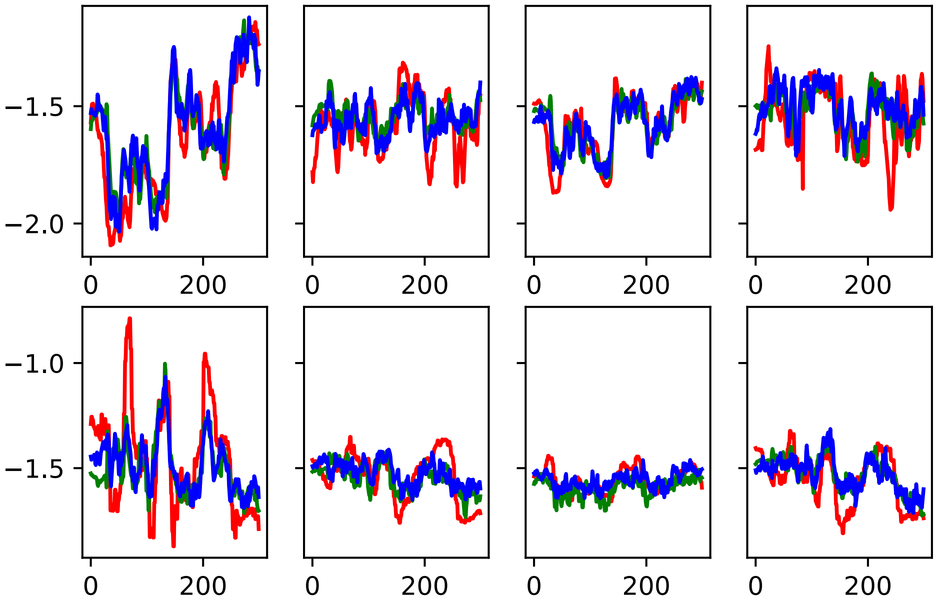
(g) BiLSTM model trained on Mandarin subset and tested on English subset.



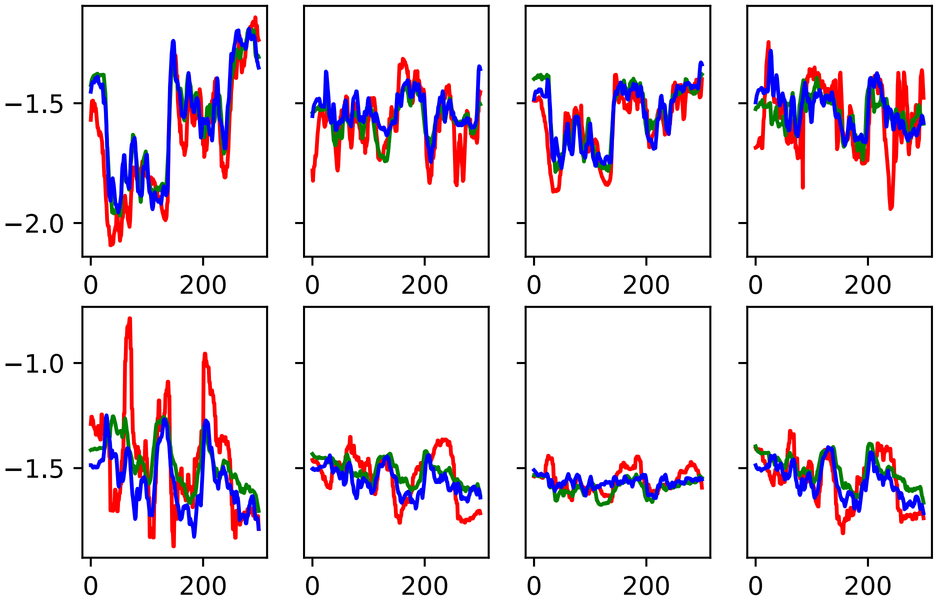
(h) CW model trained on Mandarin subset and tested on English subset.



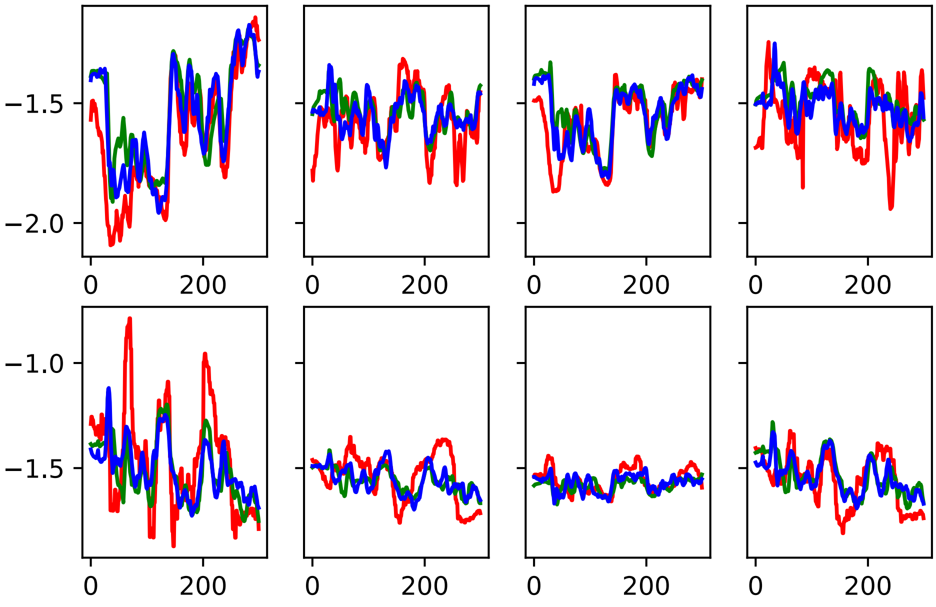
(i) LSTM model trained on Mandarin subset and tested on English subset.



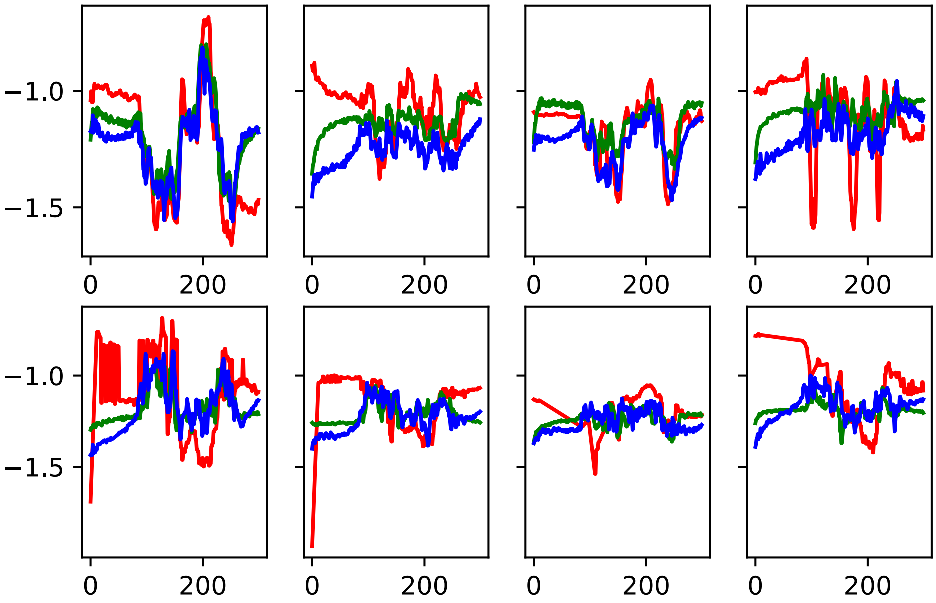
(j) BiLSTM model trained on Mandarin subset and tested on Mandarin subset.



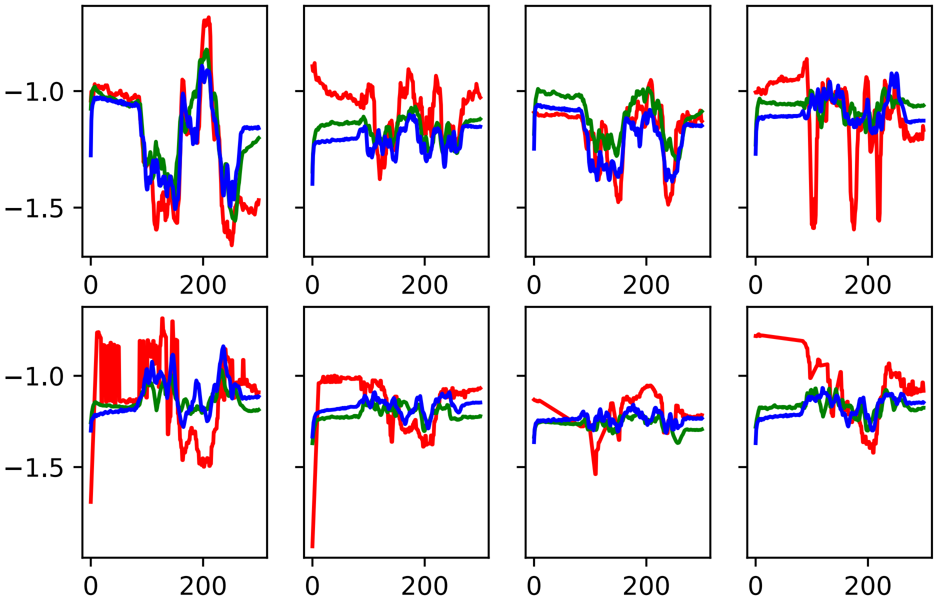
(k) CW model trained on Mandarin subset and tested on Mandarin subset.



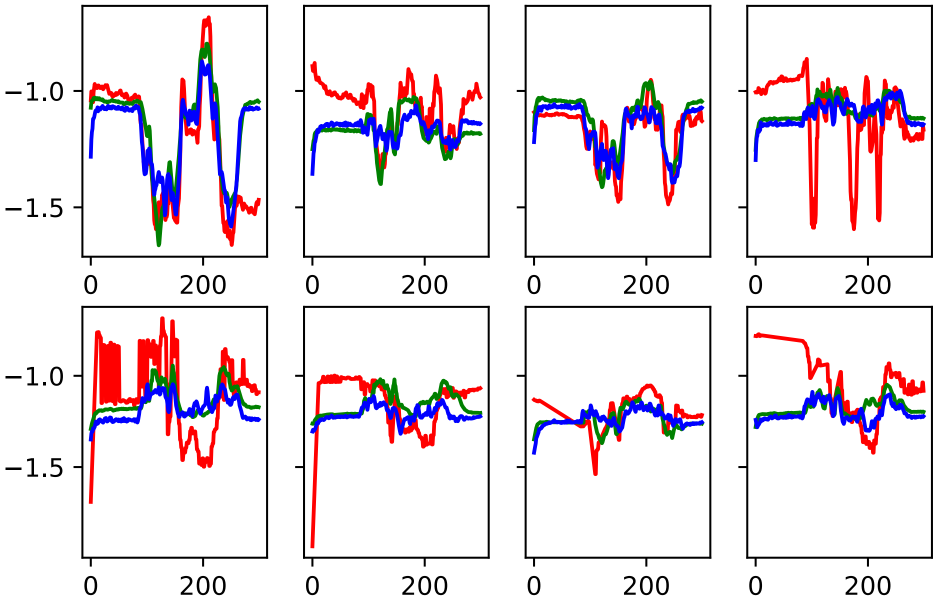
(l) LSTM model trained on Mandarin subset and tested on Mandarin subset.



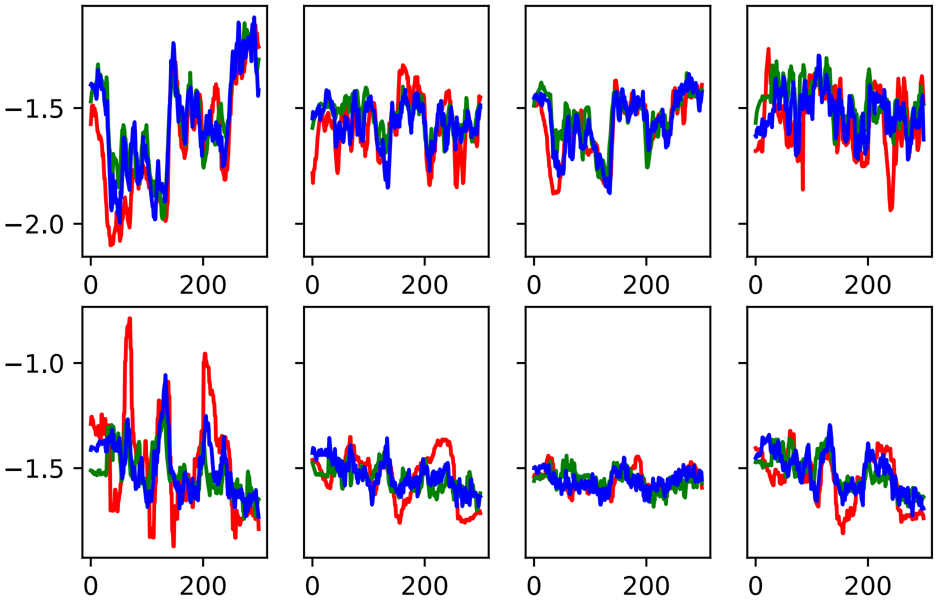
(m) BiLSTM model trained on bilingual subset and tested on English subset.



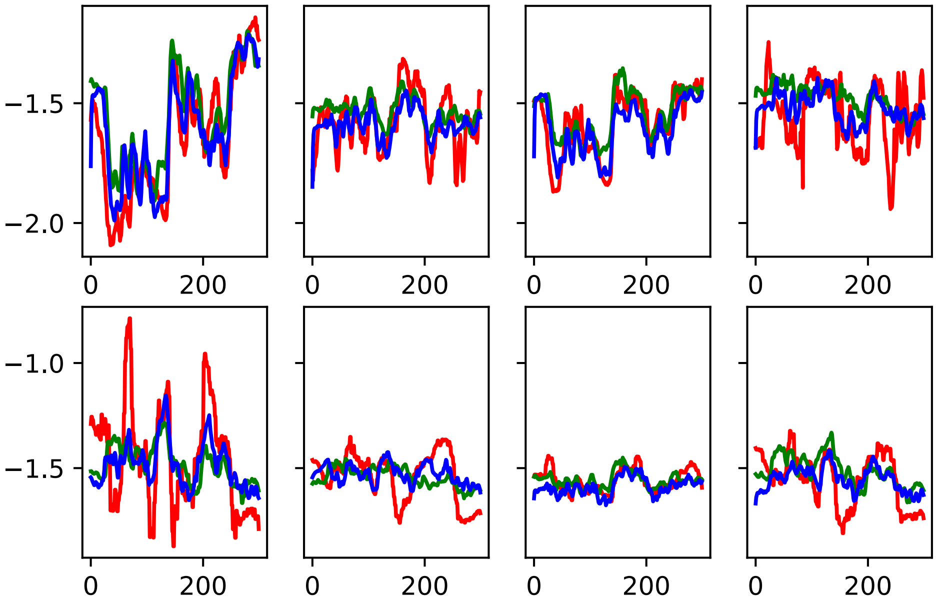
(n) CW model trained on bilingual subset and tested on English subset.



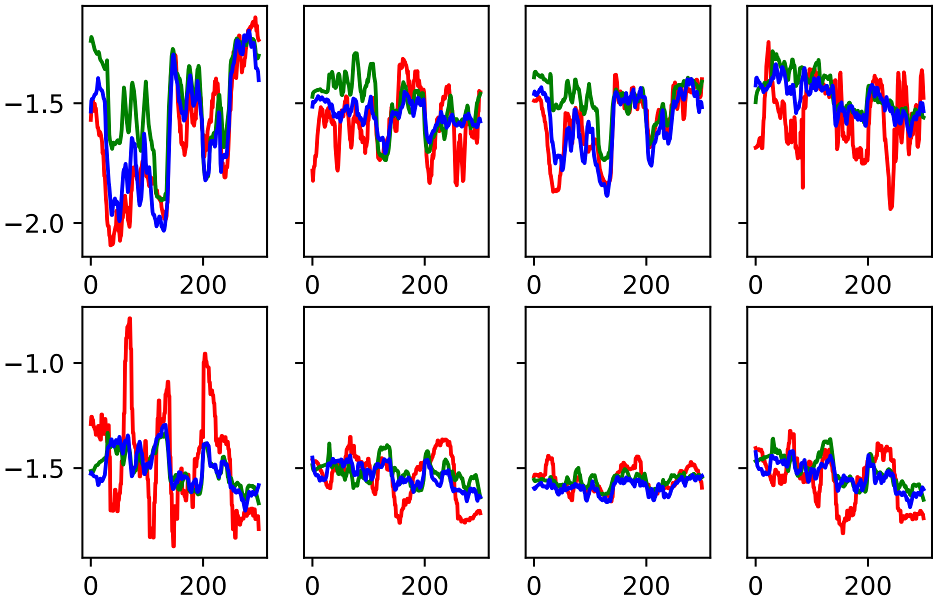
(o) LSTM model trained on bilingual subset and tested on English subset.



(p) BiLSTM model trained on bilingual subset and tested on Mandarin subset.



(q) CW model trained on bilingual subset and tested on Mandarin subset.



(r) LSTM model trained on bilingual subset and tested on Mandarin subset.

Figure 4.15: 8 articulator curves of ground-truth and predictions from different models. The red curve is the ground-truth, the predictions of models trained on MFCC features without deltas are plotted in green, and those trained on MFCC deltas are in blue.

The English testing utterance is the first 5 seconds of the 14th recording, “A roll of wire lay near the wall.” The Mandarin testing utterance is the first 5 seconds of the 20th recording, “*bao4 gao zuo4 zhe3 a1 huo4 wei1 lian2 zhi3 chu1, er4 ling2 ling2 qi1 nian2*...” The testing examples in the two languages are the same for Figure 4.6 and Figure 4.15.

# Chapter 5

# General Discussion

This paper discusses how three factors, namely acoustic representations, model types, and training language subsets, affect the performance of AAI models. We find that the first two factors do not have significant differences among them, while the language factor plays a major role. AAI models cannot map well on a language it has not seen before, so the bilingual model performs the best since it is trained on two languages.

The benefit of ultrasound is fully demonstrated in this paper. Without requiring much resources and time, ultrasound provides us with an inexpensive and quick way to record a small dataset. Considering the models' performance on ultrasound data is comparable to the same models' on EMA data, the accuracy of ultrasound is promising. Nevertheless, DLC is susceptible to the quality of the input video, especially the brightness and contrast. In this paper, we note that the lip camera video is not traced well by DLC, but we argue that the tracking points still encode their relative positional information instead of absolute coordinates.

If we consider the models' performance on tongue features and lip features independently, we find that the RMSE for tongue features is larger than the RMSE for lip features, as Table 5.1 shows. Smaller RMSEs of lip features might look counter-intuitive, as we expect the more accurate ultrasound features to have lower RMSEs. However, this can be explained by the enlargement of the videos, where the tongue and lips are not comparable since the units are in pixels instead of millimeters. The ultrasound probe is directly against the digastric muscle, while the lip camera is fixated distantly in front of the subject. A transformation to real-life measurements is needed as an amendment to this flaw. Another interpretation of small RMSE for lip features subsequently supports the fact that the DLC model suffers from tracking lip points. The pre-trained model returns a smaller lip shape, which makes the articulator movements smaller than in reality. The AAI models learn to fit better to the more stable features, therefore the RMSEs for lip features are lower than tongue features, which have more drastic changes.

Table 5.1: RMSEs and PPMCs of tongue and lip features, grouped by model types, train set, and test set. The RMSE unit is in pixels (px).

| model | train set | test set | tongue | | lips | |
|---|---|---|---|---|---|---|
| | | | RMSE | PPMC | RMSE | PPMC |
| BiLSTM | Eng | Eng | 2.864 | 0.996 | 2.349 | 0.990 |
| | Eng | Mand | 6.321 | 0.993 | 2.277 | 0.997 |
| | Mand | Eng | 9.814 | 0.985 | 4.316 | 0.987 |
| | Mand | Mand | 1.517 | 0.997 | 0.756 | 0.997 |
| | Both | Eng | 2.921 | 0.996 | 2.422 | 0.990 |
| | Both | Mand | 1.524 | 0.997 | 0.739 | 0.997 |
| CW | Eng | Eng | 2.897 | 0.996 | 2.273 | 0.991 |
| | Eng | Mand | 6.467 | 0.993 | 2.475 | 0.996 |
| | Mand | Eng | 8.349 | 0.987 | 4.771 | 0.986 |
| | Mand | Mand | 1.501 | 0.998 | 0.738 | 0.997 |
| | Both | Eng | 3.008 | 0.996 | 2.390 | 0.990 |
| | Both | Mand | 1.653 | 0.997 | 0.700 | 0.998 |
| LSTM | Eng | Eng | 2.596 | 0.997 | 2.313 | 0.990 |
| | Eng | Mand | 6.622 | 0.993 | 2.585 | 0.996 |
| | Mand | Eng | 8.867 | 0.987 | 4.712 | 0.987 |
| | Mand | Mand | 1.423 | 0.998 | 0.743 | 0.997 |
| | Both | Eng | 2.954 | 0.996 | 2.488 | 0.989 |
| | Both | Mand | 1.608 | 0.997 | 0.769 | 0.997 |

In sum, we would conclude that this pipeline is suitable for AAI tasks, but more research is needed to explore the best preprocessing procedures. We would like to propose an experiment where different ways of extracting articulatory features are compared while the AAI mapping model is controlled. Instead of using pre-trained models of tongue and lip feature tracking, DLC can be manually labeled, so if there is sufficient time and manpower, it would be helpful to see the results between pre-trained models and human-labeled ones to test whether the effort of human labeling is valuable.

## 5.1 Future work

### 5.1.1 Problems in data

Seeing the models trained on Mandarin data tend to predict lower values (subfigure (g)-(i) in Figure 4.15), we suspect that the original training data might be biased. We look at the raw data and find that Mandarin has lower values in many features, as Table 5.2 records. Only in one feature, namely the x coordinate of the first point on the tongue root, that the English dataset has a lower mean than Mandarin's. This might arise from the fact that the subject's first language is Mandarin, and he speaks Mandarin in a more relaxed way, resulting in smaller articulatory gestures. Another possible explanation is that the Mandarin recordings are recorded after the English data, resulting in a more relaxed and adjusted articulation for Mandarin.

Table 5.2: Mean ground-truth values of the 44 output features in the English and Mandarin dataset and the differences between the two. The unit of difference is in pixels (px).

| body parts | X coordinate | | | Y coordinate | | |
|---|---|---|---|---|---|---|
| | English | Mandarin | difference | English | Mandarin | difference |
| vallecula | -1.650 | -2.157 | 0.507 | -1.598 | -2.154 | 0.556 |
| tongueRoot1 | -2.536 | -2.503 | **-0.033** | -1.601 | -2.155 | 0.553 |
| tongueRoot2 | -1.552 | -1.729 | 0.177 | -1.611 | -2.171 | 0.559 |
| tongueBody1 | -1.576 | -1.763 | 0.187 | -1.600 | -2.092 | 0.493 |
| tongueBody2 | -1.524 | -1.809 | 0.284 | -1.540 | -1.887 | 0.347 |
| tongueDorsum1 | -1.541 | -1.890 | 0.349 | -1.508 | -1.802 | 0.293 |
| tongueDorsum2 | -1.561 | -1.911 | 0.350 | -1.546 | -1.988 | 0.442 |
| tongueBlade1 | -1.568 | -1.961 | 0.393 | -1.596 | -2.098 | 0.503 |
| tongueBlade2 | -1.583 | -2.015 | 0.432 | -1.560 | -2.072 | 0.512 |
| tongueTip1 | -1.588 | -2.034 | 0.446 | -1.544 | -2.004 | 0.459 |
| tongueTip2 | -3.391 | -3.514 | 0.123 | -1.540 | -1.952 | 0.412 |
| hyoid | -1.556 | -2.112 | 0.556 | -1.625 | -2.195 | 0.570 |
| mandible | -1.606 | -2.167 | 0.561 | -1.615 | -2.187 | 0.572 |
| shortTendon | -1.616 | -2.191 | 0.575 | -1.635 | -2.212 | 0.576 |
| leftLip | -1.757 | -2.001 | 0.244 | -1.958 | -3.843 | 1.885 |
| rightLip | -2.074 | -4.915 | 2.841 | -1.925 | -3.359 | 1.434 |
| topleftinner | -3.346 | -3.998 | 0.652 | -2.714 | -3.275 | 0.561 |
| bottomleftinner | -3.217 | -3.570 | 0.353 | -2.793 | -4.423 | 1.630 |
| topmidinner | -3.889 | -4.893 | 1.004 | -1.755 | -3.156 | 1.401 |
| bottommidinner | -3.954 | -4.701 | 0.748 | -2.628 | -4.235 | 1.607 |
| toprightinner | -3.853 | -4.961 | 1.108 | -2.730 | -3.508 | 0.778 |
| bottomrightinner | -3.794 | -4.912 | 1.118 | -2.947 | -4.190 | 1.242 |

Future work includes better control over subject pronunciation during recording, or mixing English and Mandarin prompt sentences similar to an “ABAB” experiment design. This allows us to balance the portions of multilingual data and remove errors caused by chronological order and experience. However, it is still possible that this phenomenon is language-inherent, and the “ABAB” recording protocol cannot get rid of this disparity. In that case, the best solution is to find a more consistent normalization across languages.

### 5.1.2 Data scarcity

The TaL corpus has multiple speaker data in TaL80, with a total recording length of 21.90 hours. Moreover, TaL1 has more than 2.1 hours of single-speaker English data. This current dataset size is much smaller, with around 0.5 hours of English and $\sim$ 1.5 hours of Mandarin speech for only one speaker. Aside from data scarcity, the imbalance between the two languages is also a problem. We intend to record more data in the future for the two languages to match.

The uneven lengths of English and Mandarin recordings are also considered a variable in the training process. The longest English recording is 695 frames, while the one in Mandarin is 2667. This makes the English recordings shorter than the Mandarin ones, and hard to balance the amount of two languages in the bilingual training dataset. In this paper, our approach to deal with this length disparity is to apply zero-padding to all recordings. We control the *number* of training data at 300 recordings, but the models learn from unequal amounts of *frames*, as the English training data is mostly padded with zeros and is removed in the model training. This is an uncontrolled variable in this paper.

To solve this data inequality from its source, the best way is to normalize the length of prompt sentences, but one compromising solution to utilize existing recordings is to repeat shorter utterances to match the longer ones. This way, the majority of data remains meaningful.

### 5.1.3 Acoustic representation

Besides MFCCs, we also consider other types of representation, including self-supervised learning (SSL) pre-trained models to represent waveforms, as [Medina et al., 2022] finds robust performance in 3D tongue modeling and [Udupa et al., 2023] finds higher performance in AAI. It would be precious to see how the state-of-the-art SSL representation can be used on this newly designed dataset. Other canonical acoustic representations include line spectral frequencies (LSF), which is used in the mngu0 dataset [Richmond et al., 2011]. Future researchers can use the same representation to compare with this well-researched dataset. The original recordings are preserved for future researchers to manipulate the representations as they wish.

### 5.1.4 Model types

The SOTA solution for AAI is models with bidirectional gated recurrent units (BiGRU) [Wu et al., 2023], [Siriwardena et al., 2022]. GRU has an advantage over LSTM models when the train set is small, and should perform better in this small dataset.

# Chapter 6

# Conclusions

In this paper, we explore how different model types and training languages affect the performance of articulatory-to-acoustic inversion mapping. Using ultrasound and camera video to capture the movements of 22 articulators on the tongue and lips, we trace articulatory features in x-y coordinates by DeepLabCut.

Three types of LSTM models (BiLSTm, CW, and LSTM) trained on two types of acoustic representations (MFCC and MFCC with deltas) of three language subsets of training data (English, Mandarin, and bilingual) predict the 44 articulatory coordinate features. We evaluate the models' performances by root mean square error (RMSE) and Pearson product-moment correlation (PPMC). We found that all models fit well with the ground-truth curve, as PPMCs are close to 0.99. RMSE results have shown that the BiLSTM model trained on bilingual data performs the best. This supports our hypothesis that AAI is language-dependent and training on bilingual data keeps its performance comparable with two independent models each trained on monolingual data.

We conclude that model types do not affect much on AAI performance, but training language plays a vital part as AAI is essentially language-dependent. We also test the feasibility of using DeepLabCut to trace articulatory features from ultrasound and lip videos, which is a more economical substitution for EMA data. Compared to previous papers, we do not test many acoustic representations, nor do we use state-of-the-art models, but we hope that this paper serves as a small step toward further research on multilingual datasets.

# Bibliography

[Articulate Instruments Ltd, 2010] Articulate Instruments Ltd (2010). *SyncBrightUp Users Manual: Revision 1.10.* Articulate Instruments Ltdg, Edinburgh, UK.

[Badin et al., 2010] Badin, P., Youssef, A. B., Bailly, G., Elisei, F., and Hueber, T. (2010). Visual articulatory feedback for phonetic correction in second language learning. In *L2SW, Workshop on" Second Language Studies: Acquisition, Learning, Education and Technology*, pages P1–10.

[Beeson and Richmond, 2023] Beeson, R. and Richmond, K. (2023). Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video. In *Interspeech 2023*. ISCA.

[Chi et al., 2021] Chi, P.-H., Chung, P.-H., Wu, T.-H., Hsieh, C.-C., Chen, Y.-H., Li, S.-W., and Lee, H.-y. (2021). Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE.

[Cho et al., 2023] Cho, C. J., Wu, P., Mohamed, A., and Anumanchipalli, G. K. (2023). Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. `https://keras.io`.

[Engwall et al., 2006] Engwall, O., Bälter, O., Öster, A.-M., and Kjellström, H. (2006). Designing the user interface of the computer-based speech training system artur based on early user tests. *Behaviour & Information Technology*, 25(4):353–365.

[Erickson et al., 2004] Erickson, D., Iwata, R., Endo, M., and Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in mandarin chinese. In *International symposium on tonal aspects of languages: With emphasis on tone languages*.

[Fairbanks, 1940] Fairbanks, G. (1940). Voice and articulation drillbook. ny.

[Garofolo, 1993] Garofolo, J. S. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

[Ghosh and Narayanan, 2010] Ghosh, P. K. and Narayanan, S. (2010). A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–2172.

[Ji et al., 2014] Ji, A., Berry, J. J., and Johnson, M. T. (2014). The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7719–7723.

[Jones, 2017] Jones, D. K. (2017). *Development of kinematic templates for automatic pronunciation assessment using acoustic-to-articulatory inversion*. PhD thesis, Marquette University.

[Kingma and Ba, 2017] Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

[Kjellström and Engwall, 2009] Kjellström, H. and Engwall, O. (2009). Audiovisual-to-articulatory inversion. *Speech Communication*, 51(3):195–209.

[Liberman and Mattingly, 1985] Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

[Ling et al., 2020] Ling, S., Liu, Y., Salazar, J., and Kirchhoff, K. (2020). Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433. IEEE.

[Liu et al., 2021] Liu, A. T., Li, S.-W., and Lee, H.-y. (2021). Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

[Liu et al., 2020] Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

[Liu et al., 2015] Liu, P., Yu, Q., Wu, Z., Kang, S., Meng, H., and Cai, L. (2015). A deep recurrent approach for acoustic-to-articulatory inversion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4450–4454. IEEE.

[Mathis et al., 2018] Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289.

[Mawass et al., 1997] Mawass, K., Badin, P., and Bailly, G. (1997). Synthesis of fricative consonants by audiovisual-to-articulatory inversion. In *Fifth European Conference on Speech Communication and Technology*.

[McFee et al., 2023] McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Halvachs, E., Thomé, C., Robert-Stöter, F., Bittner, R., Wei, Z., Weiss, A., Battenberg, E., Choi, K., Yamamoto, R., Carr, C., Metsai, A., Sullivan, S., Friesch, P., Krishnakumar, A., Hidaka, S., Kowalik, S., Keller, F., Mazur, D., Chabot-Leclerc, A., Hawthorne, C., Ramaprasad, C., Keum, M., Gomez, J., Monroe, W., Morozov, V. A., Eliasi, K., nullmightybofo, Biberstein, P., Sergin, N. D., Hennequin, R., Naktinis, R., beantowel, Kim, T., Åsen, J. P., Lim, J., Malins, A., Hereñú, D., van der Struijk, S., Nickel, L., Wu, J., Wang, Z., Gates, T., Vollrath, M., Sarroff, A., Xiao-Ming, Porter, A., Kranzler, S., Voodoohop, Gangi, M. D., Jinoz, H., Guerrero, C., Mazhar, A., toddrme2178, Baratz, Z., Kostin, A., Zhuang, X., Lo, C. T., Campr, P., Semeniuc, E., Biswal, M., Moura, S., Brossier, P., Lee, H., and Pimenta, W. (2023). librosa/librosa: 0.10.0.post2.

[Medina et al., 2022] Medina, S., Tome, D., Stoll, C., Tiede, M., Munhall, K., Hauptmann, A. G., and Matthews, I. (2022). Speech driven tongue animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20406–20416.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

[Parker et al., 2009] Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2009). Chinese gigaword fourth edition. *LDC2009T27*.

[Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE.

[Ribeiro et al., 2021] Ribeiro, M. S., Sanger, J., Zhang, J.-X., Eshky, A., Wrench, A., Richmond, K., and Renals, S. (2021). Tal: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1109–1116. IEEE.

[Richmond, 2002] Richmond, K. (2002). *Estimating articulatory parameters from the acoustic speech signal*. PhD thesis, University of Edinburgh.

[Richmond et al., 2011] Richmond, K., Hoole, P., and King, S. (2011). Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association*.

[Rothauser, 1969] Rothauser, E. (1969). Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

[Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[Shahrebabaki et al., 2019] Shahrebabaki, A. S., Olfati, N., Imran, A. S., Siniscalchi, S. M., and Svendsen, T. (2019). A phonetic-level analysis of different input features for articulatory inversion. In *Interspeech*, pages 3775–3779.

[Siriwardena et al., 2022] Siriwardena, Y. M., Sivaraman, G., and Espy-Wilson, C. (2022). Acoustic-to-articulatory speech inversion with multi-task learning. *arXiv preprint arXiv:2205.13755*.

[Taylor et al., 1998] Taylor, P., Black, A. W., and Caley, R. (1998). The architecture of the festival speech synthesis system. In *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*.

[Tiede et al., 2017] Tiede, M., Espy-Wilson, C. Y., Goldenberg, D., Mitra, V., Nam, H., and Sivaraman, G. (2017). Quantifying kinematic aspects of reduction in a contrasting rate production task. *The Journal of the Acoustical Society of America*, 141(5):3580–3580.

[Torng, 2000] Torng, P.-C. (2000). *Supralaryngeal articulator movements and laryngeal control in Mandarin Chinese tonal production*. University of Illinois at Urbana-Champaign.

[Udupa et al., 2023] Udupa, S., Siddarth, C., and Ghosh, P. K. (2023). Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

[Vroomen et al., 2004] Vroomen, J., Keetels, M., De Gelder, B., and Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive brain research*, 22(1):32–35.

[Wang et al., 2022] Wang, J., Liu, J., Zhao, L., Wang, S., Yu, R., and Liu, L. (2022). Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4808–4812. IEEE.

[Weinberger, 2015] Weinberger, S. (2015). Speech accent archive. *George Mason University*.

[Wieling et al., 2017] Wieling, M., Sivaraman, G., and Espy-Wilson, C. (2017). Analysis of acoustic-to-articulatory speech inversion across different accents and languages. In *Proceedings of INTERSPEECH*, pages 974–978.

[Wrench, 1999] Wrench, A. (1999). The mocha-timit articulatory database.

[Wrench and Balch-Tomes, 2022] Wrench, A. and Balch-Tomes, J. (2022). Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using deeplabcut. *Sensors*, 22(3):1133.

[Wrench, 2017] Wrench, A. A. (2017). Articulate assistant advanced user guide (version 2.17.02). *Edinburgh: Articulate Instruments Ltd.*

[Wu et al., 2023] Wu, P., Chen, L.-W., Cho, C. J., Watanabe, S., Goldstein, L., Black, A. W., and Anumanchipalli, G. K. (2023). Speaker-independent acoustic-to-articulatory speech inversion.

[Yamagishi et al., 2019] Yamagishi, J., Veaux, C., MacDonald, K., et al. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR).*

[Yu, 2016] Yu, L. (2016). pinyin.