Master's Thesis

Pronunciation Learning Based on Visual Articulator Movement

Mushaffa Rasyid Ridha

Supervisor: Professor HASEGAWA Shinobu

School of Information Science
Japan Advanced Institute of Science and Technology
(Master's Degree)

September, 2024

Abstract

In today's globalized world, multilingualism is essential, increasing the demand for second language (L2) learning. Learners must master vocabulary, grammar, and pronunciation for effective communication. Pronunciation is very important, as mispronunciations can easily lead to message misinterpretation. To increase the proficiency of language learners there are human-assisted pronunciation training and computer-assisted pronunciation training.

Human-assisted pronunciation training is pronunciation training that involves a professional linguist in correcting and analyzing pronunciation problems, but this approach is costly and time-consuming, also the requirement of an expert is one disadvantage of this approach. To solve this problem, Computer-Assisted Pronunciation Training (CAPT) was developed to provide a more affordable and accessible alternative for self-directed language learning.

There are several approaches used by researchers to adapt pronunciation training concepts using technology/computer: app-based system, visual simulation-based system, AI-based system, comparative phonetic-based system, and game-based system. Most of the CAPT models have the feedback in the form of technical knowledge of speech production in phonetic knowledge or only give feedback in the form of a score or text, without any correction. This kind of feedback does not show the exact reason for the error in the pronunciation of L2 learners. Several CAPTs are also able to detect phone-level errors and show the error of pronunciation according to their phonetic label, however, detecting phone-level errors alone does not necessarily result in effective feedback for learners. While learners are informed of which phonemes and words contain mistakes, they do not receive clear guidance on which articulatory movements caused these errors. To solve this issue there is a need for a CAPT system that could analyze the pronunciation using an articulator-based system.

Pronunciation is linked to how articulators move to produce sound. By understanding the human articulator movement the pronunciation could be analyzed and features could be detected. In the research about speech production mechanisms, several methods were used to scan the actual movement of the articulator. RtMRI is particularly notable for being non-invasive, free from radiation, and providing high-resolution images of vocal tract configurations. This research proposed a method by getting the articulation contour data from rtMRI and then training a model based on the paired speech data,

which will result in a model that could generate the articulator movement with input sound.

For this task, the accuracy of the paired data is critical. However, the available dataset contains inaccuracies that need to be refined before they can be utilized for training. This research also introduces a comprehensive refinement method for the rtMRI dataset, addressing contour labeling inaccuracies through a three-step process: outlier removal, FCN-based smoothing, and point-to-curve projection. These steps significantly enhance the quality of the data, as evidenced by improved contour labels that have been evaluated through subjective assessment methods.

After the pair data of sound and label is produced, the next step is to train an articulator movement generation model. This model will need to be able to output the outline of the articulator movement with only speech input. To facilitate the model training, speech features were extracted from audio. In this research, there are 3 speech features that were investigated.

Using different audio features extraction methods, phoneme human-annotated labeling, MFCC feature, and wav2v2ec 2.0 feature for articulator movement generation model training, the resulting model is capable of generating the articulator movement from wav2vec 2.0 feature extraction method but failed using the phoneme and MFCC feature extraction methods. Using wav2vec 2.0 features, the more complex model with 12x more parameters was trained and produced a more accurate model. This results in a model that is capable of generating the movement of the articulator with only speech data, using wav2vec 2.0 feature extraction.

The model's capability to detect pronunciation could be used for pronunciation training, as it is capable of generating the general trend of articulator movement. This visualization feedback then can be used to enhance the learning ability of pronunciation of L2 learner, to detect errors in their pronunciation, or to compare themselves with the right pronunciation.

Keywords: Speech-to-articulatory, articulator movement generation, real-time magnetic resonance imaging (rtMRI), automatic speech recognition (ASR), Wav2vec 2.0 pertaining, language learning, automatic pronunciation assessment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Pronunciation Training

The prevalence of human interaction among people from different nationalities, cultures, and linguistic backgrounds has increased. One significant challenge to effective international communication is the language barrier. Since language serves as the fundamental and crucial means of connecting with one another, acquiring proficiency in the spoken aspect of a new language creates opportunities to surmount barriers in cross-cultural communication.

In today's globalized world, multilingualism is essential, increasing the demand for second language (L2) learning. Learners must master vocabulary, grammar, and pronunciation for effective communication. Pronunciation is very important, as mispronunciations can easily lead to message misinterpretation. To increase the proficiency of language learners there are human-assisted pronunciation training and computer-assisted pronunciation training.

1.1.1 Human Assisted Pronunciation Training

Human-assisted pronunciation training is pronunciation training that involves a professional linguist in correcting and analyzing pronunciation problems. The knowledge of the movement of the articulator is very crucial in detecting pronunciation errors and correcting the pronunciation. Learning a language with native experts often entails the correction of articulatory movements to improve pronunciation accuracy. However, professional human-assisted training is often costly and time-consuming. Then researchers start integrating the advancement in technology including automatic error detection or social networking services to reduce the cost and increase the efficiency of the learning. [14]

1.1.2 Computer Assisted Pronunciation Training

The integration of technology in many human-assisted tasks also produces development in the form of Computer-assisted Pronunciation Training (CAPT). Many CAPT [14, 15, 16] applications have been developed to aid L2 learners, providing a more affordable and accessible alternative for self-directed language learning. Generally, CAPT model is divided into two tasks: automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). APA primarily evaluates speech to assign pronunciation proficiency scores that closely match those given by human evaluators, addressing aspects like accentedness, fluency, and comprehensibility across different granularities such as phones, words, and sentences [17, 18, 19, 20, 4, 11]. MDD focuses on identifying pronunciation errors and providing instant feedback to help L2 learners improve their speaking skills [21, 22, 23, 24, 19]. Studies have analyzed the most frequent phonetic errors and explored distinctive features and classifiers. Additionally, some systems incorporate automatic speech recognition (ASR) technology, integrating possible errors into the lexicon to provide information about pronunciation mistakes [19, 4].

1.2 Speech and Articulatory Movement

Speech production mechanisms are defined at a peripheral level by the acoustic outputs they generate and the dynamic movements of the vocal tract over time. As a result, substantial research has focused on gathering both acoustic and articulatory data to better understand and analyze the sounds of the world's languages [25]. This data has supported the development of articulatory models of speech production by researchers [26, 27, 28, 29, 30] with the aim to infer articulatory speech motor control schemes, while methods for modeling relationships between gestures and sounds using statistical approaches have also been explored [31, 32]. Additional research has investigated the overall geometry of the vocal tract and its relation to vocal tract acoustics [33, 34]. Further, focused studies on individual articulators have contributed to developing biomechanical articulatory models [35, 36] and examining speech temporal coordination, as seen in studies by [37, 38].

To explore directly the articulator movement various visualization techniques have been employed, including X-ray imaging [39, 40] Electromagnetic Articulography (EMA) [41, 1, 11] and real-time Magnetic Resonance Imaging (rtMRI). EMA is a method to measure the position of parts of an articulator using several placed coils in the tongue and other parts of the mouth, this method allows observation of movement in several areas of the mouth. Using

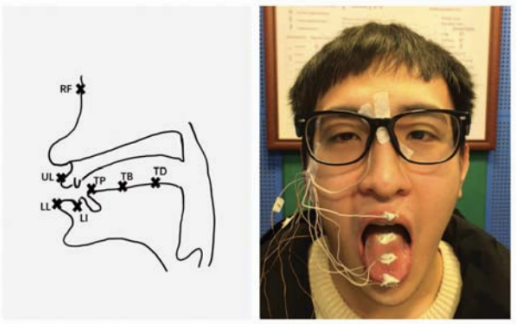


Figure 1.1: Left: EMA sensor's position, Right: Setup of EMA recording. [1]

the EMA method gives 6 xy-points data, that could observe areas between the tongue and palate and two lips. MRI data shows the image scan of including glottal, uvular, soft, and hard palate, tongue, and lips.

1.3 Limitation and Challenge

This section will explain the limitations and challenges of the existing methods in relation to CAPT with articulator movement.

1.3.1 Feedback Technique

Several approaches was used for pronunciation feedback training: app-based system, visual simulation-based system, AI-based system, comparative phonetic-based system, and game-based system [14]. The app-based system was the most used system based on the literature study by Chen [14], this tool uses using available mobile app or Social Networking Sites (SNSs) to give instruction and and pronunciation training [12]. The visual simulation-based system using the spectrogram image including the pitch change intonation change [17, 3]. For AI-based systems, different research tried using different kinds of models to produce a model that could detect the error in pronunciation [4, 20]. The comparative phonetic-based approach is a system that compares the phonetic features [5], usually used by an expert who is fluent in their native languages. In this method, the phonemes of a learner's native language are compared to those in English through stochastic (probabilistic) analysis. The Computer-Assisted Pronunciation Training (CAPT) system records the learners' speech, identifies mispronunciations, and provides improvement suggestions based on their native languages. Game-based approaches employ interactive, goal-oriented tasks in both formal and informal settings to

teach pronunciation. Such a system can simulate real-world conversations and train learners to speak appropriately in various contexts [6].

As explained before in the section 1.1.2, there are 2 types of CAPT, the ones that focus on automatic pronunciation assessment (APA) and the ones that focus on mispronunciation detection and diagnosis (MDD). From different tools for pronunciation feedback, visual feedback was found to be effective in the L2 classroom for the teaching of segmental features of pronunciation. Using visual feedback of the difference between spectrograms of speech, L2 learners achieve significant improvement, more native-like, productions of the pronunciation [42]. To achieve a system that could give both APA and MDD objectives at the same time, there is a need for system development that could assess pronunciation but also be used for diagnosis tasks. A CAPT system that utilizes an articulator-based approach with articulatory movement would provide clear assessment and precise feedback diagnosis by visually demonstrating the physical interactions of the articulators during speech production.

1.3.2 Articulator Movement Extraction

To be able to train the model that could represent the movement of the articulator to analyze the pronunciation, the dataset that has the pair data for articulator movement and its corresponding speech sound is crucial. Among the 3 physical feature extraction from the actual movement of the articulator, rtMRI is particularly notable for being non-invasive, free from radiation, and providing high-resolution images of vocal tract configurations. This makes rtMRI an excellent tool for examining the dynamic movements of the tongue, lips, and palate during speech, including the unseen areas such as tongue radical to pharyngeal, the uvular to pharyngeal, epiglottis and glottis movement, enhancing our understanding of the mechanisms underlying speech production, compared to previous research in speech-to-articulatory movement based on the EMA dataset [43] that only cover several part of the tongue. Inversely, using the MRI data it is able to generate more comprehensive speech data from articulator [9] showing the inverse task more probable. However, the available dataset for the pair data of MRI video and speech data for CAPT was found to still contain occasional errors.

1.3.3 CAPT by articulatory features

A significant step towards articulator-based pronunciation correction is the DNN articulatory model system by Duan et. all [19]. Instead of providing typical error feedback such as "You pronounced the 'th' sound incorrectly"

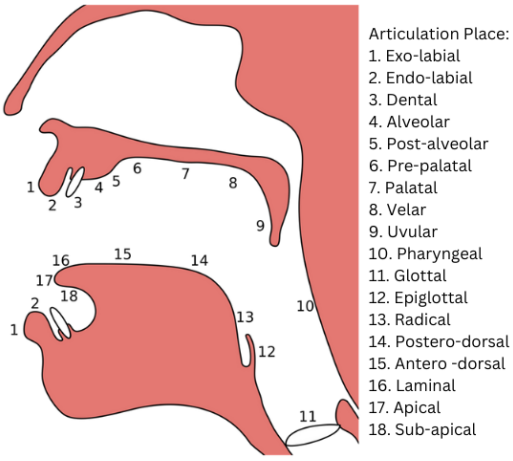


Figure 1.2: Places of articulation (passive & active). [2]

when a user pronounces the word "think" as "sink," it aims to provide more detailed guidance like "Place your tongue between your teeth and blow air softly." This system uses large corpora from Japanese and Chinese languages to model inter-language phonemes using IPA categorization, requiring a predefined phoneme set specific to Chinese and Japanese. Unfortunately, this research did not include visualization, making it difficult for learners to understand the exact position of their articulators. Therefore, a more in-depth analysis of dynamic movements, such as those of the tongue, lips, and palate during speech production, is essential for understanding the articulatory-sound relationship and providing more effective feedback.

1.4 Objective and Originality of this Thesis

To address the gap that is presented in section 1.3, our research aim is to develop an articulatory movement generation model that could be used in the CAPT system that offers valuable feedback on detailed visual articulatory movements.

1.4.1 Objective

Motivated by the principles underlying human sound production and its potential to aid pronunciation training, this research aims to develop an articulator movement generation that could be used for a Computer-Assisted Pronunciation Training system. The objectives of this research are:

- To detect articulatory movements corresponding to the recorded speech

sound, including choosing the feature extraction methods, and dataset refinement.

- To assess the effectiveness of the articulator movement generator for the pronunciation detection task.

1.4.2 Originality

Throughout our study, we introduce our novelty and contribution in the field, which are as follows:

- We proposed a necessary data refinement method to develop a refined dataset pairing speech sounds with MRI-derived articulatory contours, enabling the detection of articulatory movements.
- We propose a method to generate articulator movement generation, which can represent the movement of the articulator based on the input sound, in response to audio input.
- We evaluate the articulator movement generator's effectiveness in detecting the pronunciation.

1.5 Organization of Thesis

This thesis is structured with five chapters. The following is a quick summary of the contents of each chapter:

- Chapter 1 gives the introduction and background of this thesis. Section 1.1 gives the definition and introduction of pronunciation training including briefly current technologies of the CAPT system. Section 1.2 briefly explains the existing method for speech and articulatory relation. Section 1.3 shows the limitations and challenges of the current research. Section 1.4 demonstrates the objectives as well as the originalities of this thesis. Finally, the last section 1.5 details the organization of this thesis.
- Chapter 2 introduces CAPT systems and their approach. Section 2.1 tells briefly the focus of the chapter. Section 2.2 shows examples of app-based approaches used by the CAPT system. Section 2.3 shows examples of visual simulation-based approaches used by the CAPT system. Section 2.4 shows examples of AI-based approaches used by the CAPT

system. Section 2.5 shows examples of comparative phonetic-based approaches used by the CAPT system. Section 2.6 shows examples of game-based approaches used by the CAPT system.

- Chapter 3 details the proposed refinement method used in this research. Section 3.1 gives a brief overview of the proposed model for data refinement. Section 3.2 gives a detailed explanation of the existing articulator movement dataset. Section 3.3 explains the existing work related to USC-TIMIT and its refinement. Section 3.4 explains the proposed approach used to refine the dataset. Section 3.5 explain the experimental setup used inthis study fo data refinement step. Section 3.6 shows the result of result of the experiment. Section 3.7 gives a summary of the chapter 3.

- Chapter 4 details the proposed articulator generator method used in this study. Section 4.1 gives a brief overview of the proposed method for articulator movement generator method. Section 4.2 explains the related works related to speech and movement of articulator with rtMRI dataset. Section 4.3 explains the proposed system for articulator movement generator. Section 4.4 explains the setup used in the experiment. Section 4.5 explains the result of the experiment. 4.6 gives a summary of the chapter 4.

- Chapter 5 explain the conclusion of the research and the future direction of the research. Section 5.1 explains the conclusion of this study. 5.2 explains the future direction of this study.

Chapter 2

Literature Review

2.1 Overview

This chapter introduces CAPT systems and their approach to pronunciation training. There are five example approaches used by the CAPT system that will be explained in this chapter: App-based system, visual simulation-based system, AI-based system, comparative phonetic-based system, and game-based system.

2.2 App-based system

2.2.1 Overview

Fouz-Gonzales [12] proposed Twitter-based pronunciation instruction, sending the participants a daily tweet related to pronunciation instruction. On weekdays learners will get one tweet per day for the training phase in 22 days. There 2 groups, the experimental group that were sent tweets for the target pronunciation and the another group that were sent not pronunciation tips but different English tips (grammar tips, or contents related to them).

This study conducted a test before and after training and a post-test after a month to measure whether the participant retained the ability over time. With the result of improvement rate between 22%-30%. The result after a month, half of the participant still retain the improvement. Participants' responses after the experiment are shown in Table 2.1.

Table 2.1: The response of participant, with the standard variation denoted inside the parentheses [12]

Item	G1 n = 24	G2 n = 44	All N = 68
1. Twitter is potentially useful for educational purposes	3.6 (0.8)	3.4 (1)	3.5 (0.9)
2. Twitter is potentially useful to teach grammar	3.6 (0.8)	3.1 (1)	3.3 (0.9)
3. Twitter is potentially useful to teach false friends	4.2 (0.8)	3.6 (1.1)	3.8 (1)
4. Twitter is potentially useful to teach pronunciation	3.1 (0.9)	3.3 (1.3)	3.2 (1.1)
5. Twitter is potentially useful to teach vocabulary	4.2 (0.6)	3.9 (1.1)	4 (0.9)
6. Because of having received the tips through Twitter, I will be able to remember them better	4 (0.7)	3.5 (1.2)	3.7 (1)
7. The tips I received through Twitter are useful	4.2 (0.7)	4.2 (0.8)	4.2 (0.7)

2.2.2 Limitation

This research shows that multiple corrections or feedback using technology as a bridge without direct contact is possible and the pronunciation ability is still retained even after a period of time passed. However, there is no feedback mechanism for detecting and correcting the error.

2.3 Visual simulation-based system

2.3.1 Overview

Liu et al. [3] research the effect of CAPT system under combined 4 conditions, between scripted text and unscripted text, and between active engagement and passive engagement. Using the native English speaker recorded sound as comparison on in the training. The visual cues used here are the waveform and spectrograms. The target of the training is the suprasegmental phonology (intonation).

The results indicate that intonational gain was better using scripted than unscripted speech. This shows that there is an increase in difficulty for improvement in suprasegmental phonology (intonation) in unscripted (unplanned) speech compared to scripted (planned) speech. The findings suggest that increase in engagement and following the scripted procedure is needed to address the suprasegmental phonology in unscripted speech.

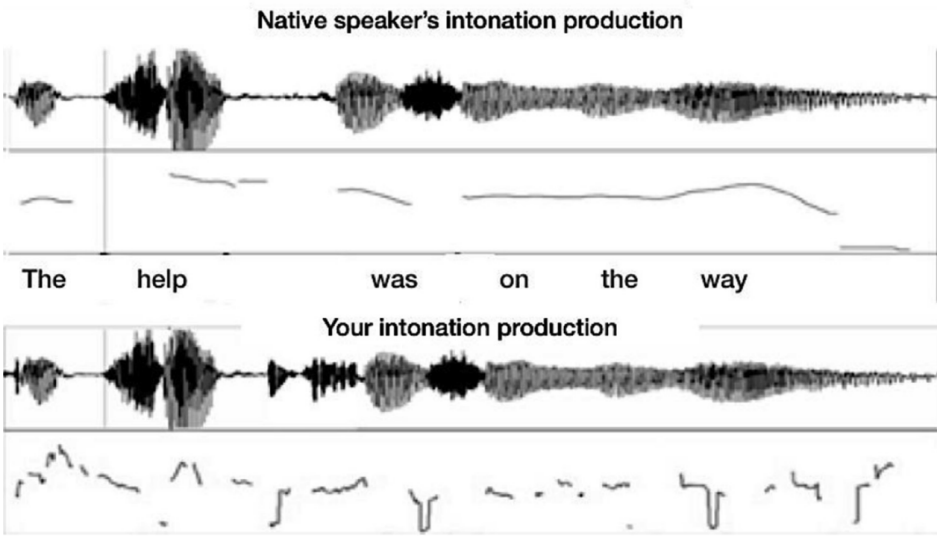


Figure 2.1: Comparison between native and learner intonational production. [3]

2.3.2 Limitation

This research shows that the CAPT system with visual feedback helps in segmenting the error and correcting the intonation from the scripted speech, but this research only addresses the suprasegmental phonology in their visual feedback and does not yet address the visual cues in the segmental phonology.

2.4 AI-based system

2.4.1 Overview

Kim et al. [4] proposed a method utilizing the SSL model (wav2vec 2.0 and HuBERT) for learning pronunciation-relevant latent representations, illustrated in Figure 2.2. There are 3 steps:

- Fine-tuning the pre-trained models with the chosen datasets.
- Extracting the contextual output of the models.
- decoding the pronunciation score from contextual information, using another layer of neural network.

Table 2.2 shows the result of pre-trained models and fine-tuned models of wav2vec 2.0, and HuBERT [44]. First, both models pre-trained models

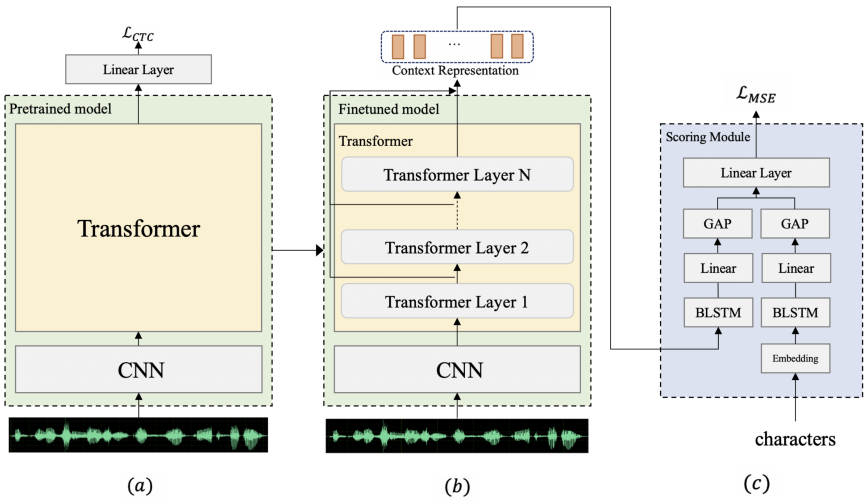


Figure 2.2: Automatic pronunciation assessment proposed by Kim's [4] based on SSL models.

shows improvements compared to baseline model, as shown in Table 2.2. The results of higher performance with wav2vec2 robust pre-trained model [45] and Large model shows that the model that was trained using real-world scenarios or larger model improve the performance compared to pre-trained counterparts.

2.4.2 Limitation

This research shows that the ASR models can detect fluency, holistic, and prosodic measures of speech using scores but not detect the exact error in pronunciation.

2.5 Comparative phonetic-based system

2.5.1 Overview

Qian et al. [5] proposed a sophisticated two-pass framework for mispronunciation detection and diagnosis (MD&D) in pronunciation detection. This method was proposed to address the typical comparative mispronunciation detection systems that rely on predefined models that predict how errors are likely to occur based on known error patterns. The system does not assume any specific types of errors beforehand (i.e., it does not use a predefined list of possible mistakes learners might make). This approach is intended to

Table 2.2: The result of assessment of models using KESL and Speechocesn762 dataset. [4]

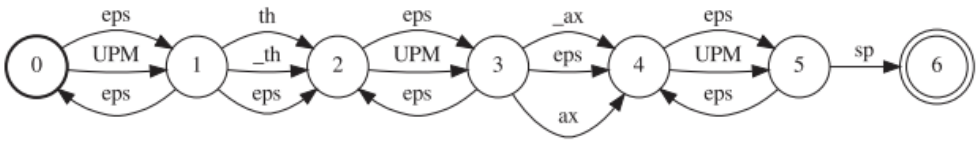
Method	KESL	Speechocean762	
	Holistic	Fluency	Prosodic
GOP	0.63	0.65	0.64
Agg + Seq	0.55	0.51	0.59
Agg + Seq + GOP	0.64	0.67	0.66
pre-trained			
wav2vec2 Base	0.65	0.72	0.72
wav2vec2 Large	0.71	0.72	0.72
wav2vec2 Robust	0.76	0.73	0.73
HuBERT Base	0.69	0.72	0.71
HuBERT Large	0.75	0.75	0.74
Finetuned			
wav2vec2 Base	0.68	0.73	0.72
wav2vec2 Large	0.78	0.73	0.72
wav2vec2 Robust	0.79	0.75	0.74
HuBERT Base	0.75	0.74	0.73
HuBERT Large	**0.82**	**0.78**	**0.77**

cover a wider range of possible errors. By not restricting the model to known error patterns, the system theoretically can identify any type of pronunciation error, thereby providing comprehensive error detection. The problem with this approach is without predefined error patterns, the system must consider a vast number of potential errors, leading to a very large set of possibilities (search space) that the system needs to evaluate. This results, the search space is so large that it becomes computationally difficult (intractable) to manage effectively. The research seeks to minimize the search space by pairing each canonical phone[1] (derived from the text prompt) with an anti-phone, which encompasses the complementary acoustic space. The anti-phone is a concept which represents sounds that are the acoustically opposite or complementary for each canonical phone. anti-phone was introduced in this research in the form of Gaussian Mixture Model (GMM)-Hidden Markov Model (HMMs) model to detect phone substitution.

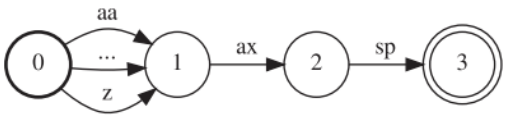
In the initial pass of recognition within this network, phonetic substitutions are detected. Integrating the filler mode[2] adds the capability of the

[1]A canonical phone is a typical or most common realization of a phoneme in a given linguistic context. It is a specific phonetic instance of a phoneme that serves as a standard or reference point within a particular discussion or analysis.

[2]Filler model was introduced in this research with the name universal phone model (UPM) which covers all the non-silence phones.



(a)



(b)

Figure 2.3: (a) The example architecture of the first pass for the word "THE". Anti-phone has the prefix '_'. 'eps' means a non-emitting skip. UPM means Universal phone model or filler model. (b) The example architecture of the second pass, when the first phone [th] was not detected. It does not appear between state 0 and 1.[5]

network to detect insertions, and deletions which allows phone skips. Next, free-phone recognition was conducted on the insertions and substitution segments to find the actual phones. Discriminative training on the two-pass framework with reduces the PER from 27.7% to 16.5%.

2.5.2 Limitation

The approach of this research is on error detection, but this method does not have visual feedback of the error that happened, and only addresses the phone notation, which would require additional knowledge of phonetic notation to understand the system feedback.

2.6 Game-based system

2.6.1 Overview

Satria et al. [6] used the game-based system CAPT to train kid on age between 8-13 years old with the proposed game called "Hello Animal".

All operations in the game use speech commands. From the start screen, user needs to navigate through the game with the command shown on the screen, either to start the game or another menu. When the game is started, the animal data will be loaded. If the user want to pause the game, they could do so at any time by saying "pause.". The user can exit the program

at any time, with a prompt asking for confirmation.

Figure 2.4: Screenshot of the game 'Hello Animal' [6]

The results of the test, based on the survey, the evaluation score of the participants was high with average 4.38 in the scale 0-5, based on user interaction score, feedback and pedagogy effect of the game.

2.6.2 Limitation

This research focused on increasing the engagement and proactive action from the participants, as seen in the subjective feedback Table **??**. However, this approach does not address much of the correct feedback for correcting the wrongness in pronunciation.

Chapter 3

Proposed rtMRI Articulatory Movement Dataset Refinement

3.1 Overview

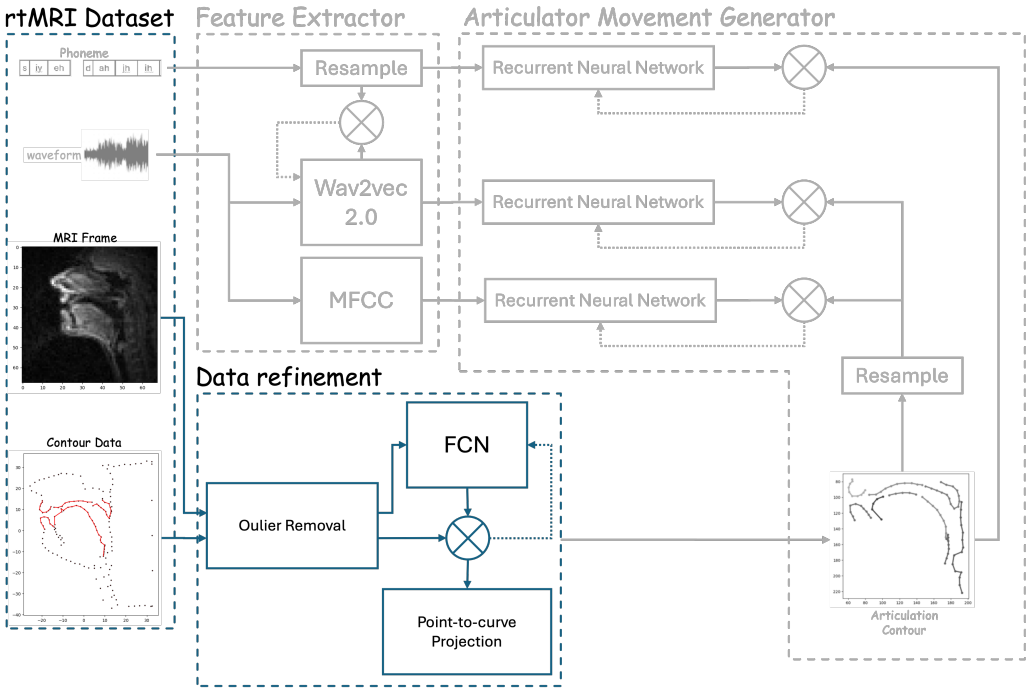


Figure 3.1: Data Refinement Section of Proposed Method

RtMRI is a method to capture the unseen area of inside the human body that is radiation-free, compared to X-ray, a non-invasive as it uses

magnetic field, and high-resolution that could be used for the visualization of vocal tract. This method provides information that could be covered by EMA (head, jaw, labial and lingual motion) and also not covered by EMA (velum, pharynx, and larynx)[13]. USC-TIMIT database [13] is a collection of resource data that is available publicly, on the website http://sail.usc.edu/span/usc-timit. This dataset provides the pair data of speech and its rtMRI data. But this dataset contains the occasional error for the contour data. A refinement method was proposed to refine the contour data, as seen in the data refinement section in the proposed method, Figure 3.1. This section will give a comprehensive explanation of the USC-TIMIT database, including the proposed refinement method.

3.2 Existing Articulatory Movement Dataset

Table 3.1: Participants detail of USC-TIMIT database. [13]

ID	Gender	Age	Birthplace
M1	Male	29	Buffalo, NY
M2	Male	33	Ann Arbor, MI
M3	Male	26	Madison, WI
M4	Male	26	St. Louis, MO
M5	Male	27	Mammoth, CA
F1	Female	23	Commack, NY
F2	Female	32	Westfield, IN
F3	Female	20	Palos Verdes, CA
F4	Female	46	Pittsburgh, PA
F5	Female	25	Brawley, CA

USC-TIMIT is an rtMRI dataset acquired by Narayanan et al. [13] It results of video with the resolution 68 $\times$ 68 pixels (2.9 $\times$ 2.9 mm) of sagital plane with the frame rate of 23.18 frames/s and 20kHz frequency for simultaneously recorded audio. The speech is the utterance of 460-sentence phonetically balanced dataset referenced in the MOCHA-TIMIT corpus [46] with 10 native speakers (5 male, 5 female) of American English with the detailed information provided in Table 3.1. Given the field of view (FOV) of 200 $\times$ 200 mm and an image resolution of 68 $\times$ 68 pixels, each pixel corresponds to approximately 2.9 mm.

This dataset also includes vocal tract contour labels using through spatial

Table 3.2: Distribution of different format points data.

Format Points	**# Frames (total)**	**# Frames (each subject)**
178	15205	15205 (F1)
180	51252	16882 (F1)
		34370 (F5)
181	28766	28766 (M3)

frequency domain-based segmentation [47]. This vocal tract contour data is only available for 3 participants, F1, F5, and M3, with 3 different formats of points that make up the contours. There are 178-points, 180-points and 181-points. There are 95,223 video frames that has the vocal tract data with the distribution of data can be seen in the table Table 3.2

3.3 Related Works on Data Refinement.

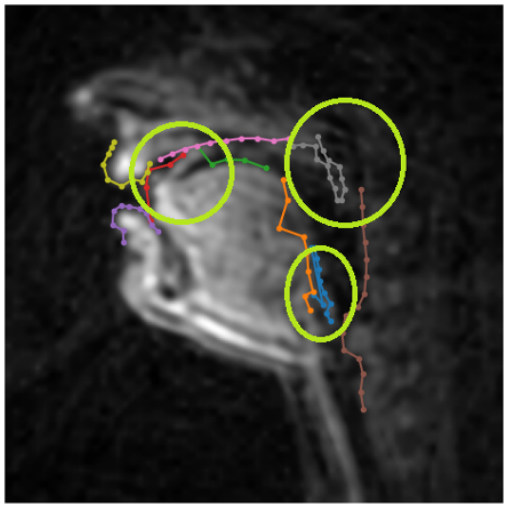


Figure 3.2: Occasional error included in the USC-TIMIT dataset

Getting the articulation movement information could be achieved by getting the dynamic movement of articulator contour, including tongue, palate, and lips. Those feature could be seen in rtMRI data and need to be captured, but the number of videos and the length of the video frame make it impossible to manually labelling the contour data by expert frame by frame, as it

is very time consuming and high cost. There are 2 automatic approach to estimate the vocal tract contour: landmark-based and segmentation-based.

Landmark-based approach utilizes methods like the active control model (ACM) [48], active shape model (ASM) [7], and articulatory-specific multiple linear regression (MLR) [49] to locate anatomical landmarks and distinguish tissue from the airway. On the other hand, segmentation-based approach is using segmentation on pixel-level or to assign tissue, then use the required segmentation the get the vocal tract contour. The labeled articulation contour for this dataset, publicly available, was produced from the work of Bresch and Narayanan [47] using unsupervised region segmentation. However, because it used unsupervised segmentation, this dataset occasionally contains errors, as demonstrated in Figure 3.2.

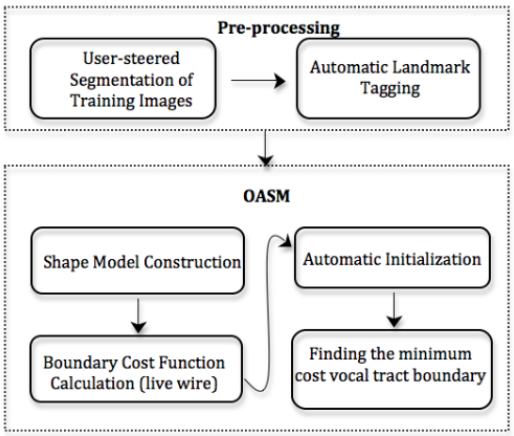


Figure 3.3: Many steps of vocal tract image segmentation in flowchart for OASMs. [7]

3.3.1 Landmark Based Approach

Raeesy et al. [7] proposed OASM method in extracting the landmark information from rtMRI data. It combines two approaches, active shape models (ASMs) [50] and live wire [51]. There are 2 steps for OASMs:

1. training and shape model construction.

2. segmentation based on the defined shape model.

Figure 3.3 shows step to contruct OASMs. Figure 3.4 shows the result of OASMs. The boundary of vocal track was tracked generally with slight error. However, the output landmark has a very high precision error as seen

in the figure and could output the error if the data was not included in the training set.

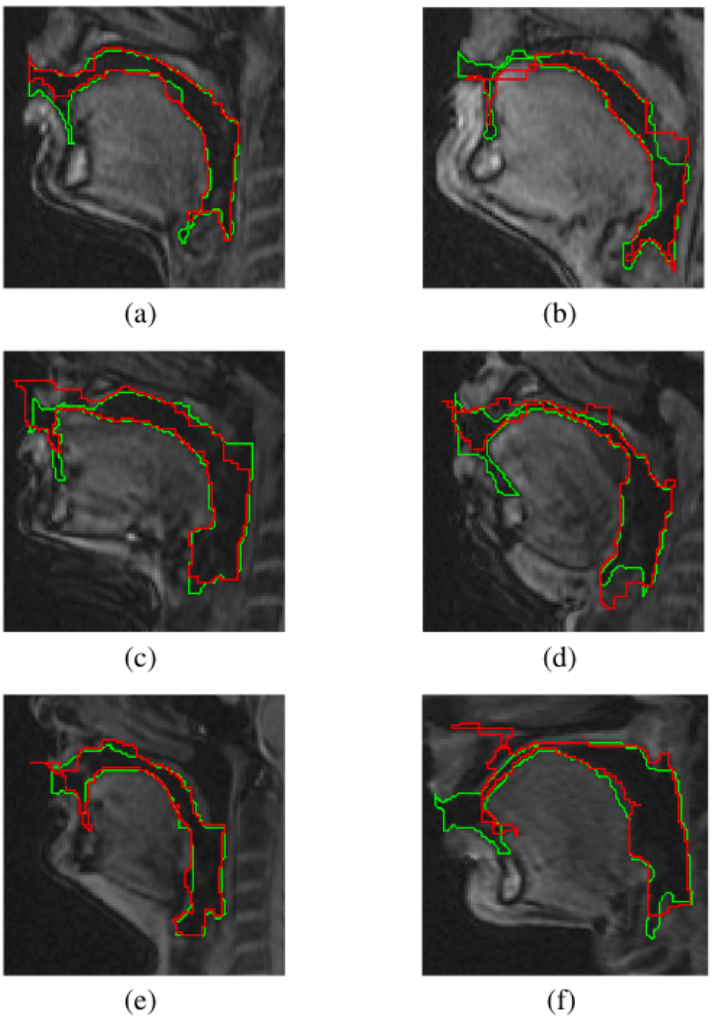


Figure 3.4: OASM result (red) with user-defined segmentation (green). With (a)-(e) were trained on other training set, and speaker (f) was not. [7]

As seen in the result image that the model not able to accurately captured the vocal track and the inaccuracy happened in many places. This method cannot be used to extract the contour data for tracking articulator movement related to pronunciation.

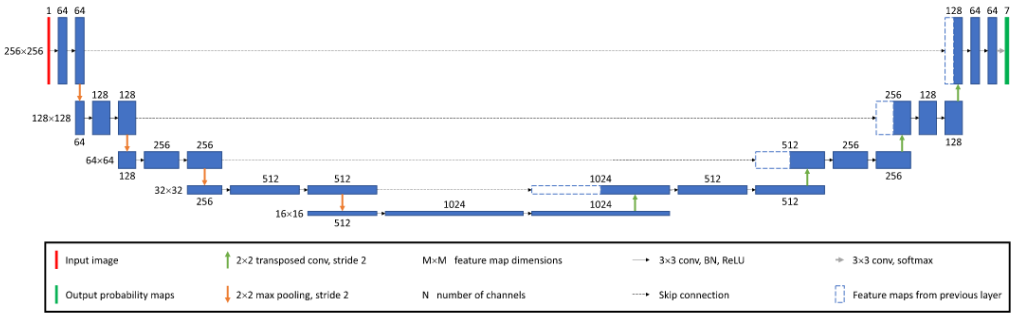


Figure 3.5: Ruthven Network architecture. Conv: convolution, ReLU: rectified linear unit, BN: batch normalisation. [8]

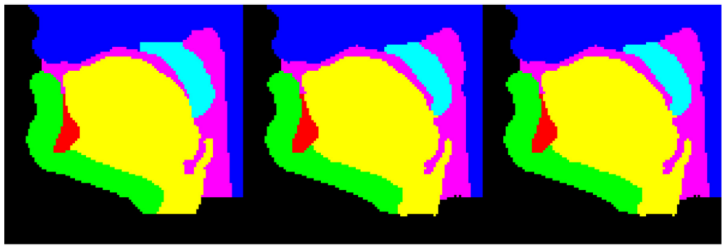


Figure 3.6: Video frame showing the segmentations, with ground truth on the left, and predicted model on the center and right. Right image is after post-processing. [8]

3.3.2 Segmentation-based Approach

Ruthven et al. [8] proposed a of Fully convolutional network inspired with U-Net model [52] to be trained with the human-annotated label. Each MRI image was labeled Physicists. Six classes was segmented by physicists: top-half of the head (hard palate and the upper lip), lower-half of the head (jaw and lower lip), tongue and epiglottis, soft palate, vocal tract and tooth space (lower incisor).

The result of the trained model can be seen in the Figure 3.6. The model is able to identify and segment the contour based on the video frame input.

3.4 Proposed Approach

Several researchers tried to assign the label for the dataset such as [7] and [8]. For method used in [8] cannot be applied to the USC-TIMIT data because the data they used is in higher precision but not available publicly, the same problem also happened using the method by Silva (2015) [53] as they also

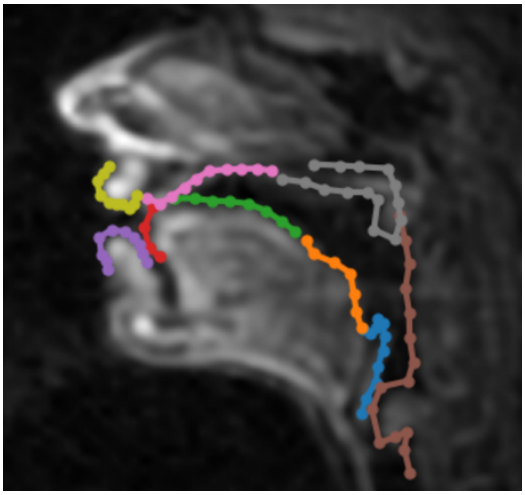


Figure 3.7: Nine areas of articulator

used an independent dataset.

The objective of the dataset refinement for this research is to be able to provide the movement of the articulator pair with the speech signal, including the publicity of the refined dataset. Having the pair data of speech signal and articulator movement is the necessary step in producing the articulator movement detection model. To produce a more high-quality dataset compared to the original dataset we propose a refinement method that would be able to produce a better pair of data of speech data and articulator movement, by tracking the contour movement of the articulator. The contour will be divided into nine areas as seen in Figure 3.7: yellow (upper lip), purple (bottom lip), pink (hard palate), red (edge tongue), green (middle tongue), orange (back tongue), blue (epiglottis), grey (uvular), brown (pharyngeal wall). To produce a more high-quality dataset compared to the original dataset. There are 3 steps for this data preparation: Outlier removal, smoothing, and point-to-curve projection.

3.4.1 Outlier Removal

For outlier removal, we are using threshold values in determining the outliers, 3.8c,d. There are several features that we use, including the size area of the articulator for the articulators that do not change the area most of the time, like the uvular and both lips. We begin by calculating the average size of each area. Next, for each dataset, we compare the size of each area to this average. Data is removed when the area size deviates significantly from the average, based on a fixed threshold value. For the articulator that is moving

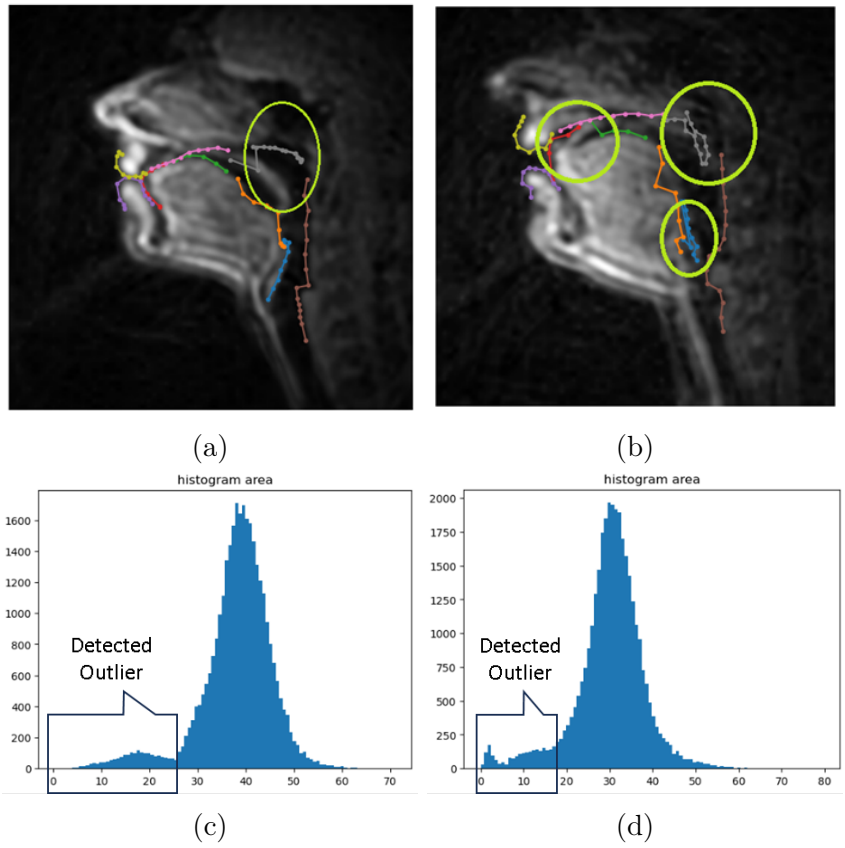


Figure 3.8: Examples of: (a) Outlier detected with articulator area size for uvular. (b) Outlier detected with the distance between contour points and their neighbor. (c,d) Area threshold for detecting the outlier

such as the tongue, the outlier will be detected using the temporal movement of the contour points and its relative distance to its neighbor points, 3.8b. For example, for uvular area in F1, we exclude data where the size is <400 square pixels and for F5 is <300 square pixels. Outlier in F1 was less compared to F5 because the image area of uvular in F5 are more blurry and smaller. Respectively, we found 248 and 2,757 outlier from each subject.

This step was conducted as when the FCN model training without the outlier removal, the model regenerates the incorrect contour as seen in 3.8a.

3.4.2 Neural Network-based Smoothing

Using neural networks to generalize the data is a common practice. Using the neural network model can generalize the tolerable error in the dataset and then as a result it removes the error from the neural network output.

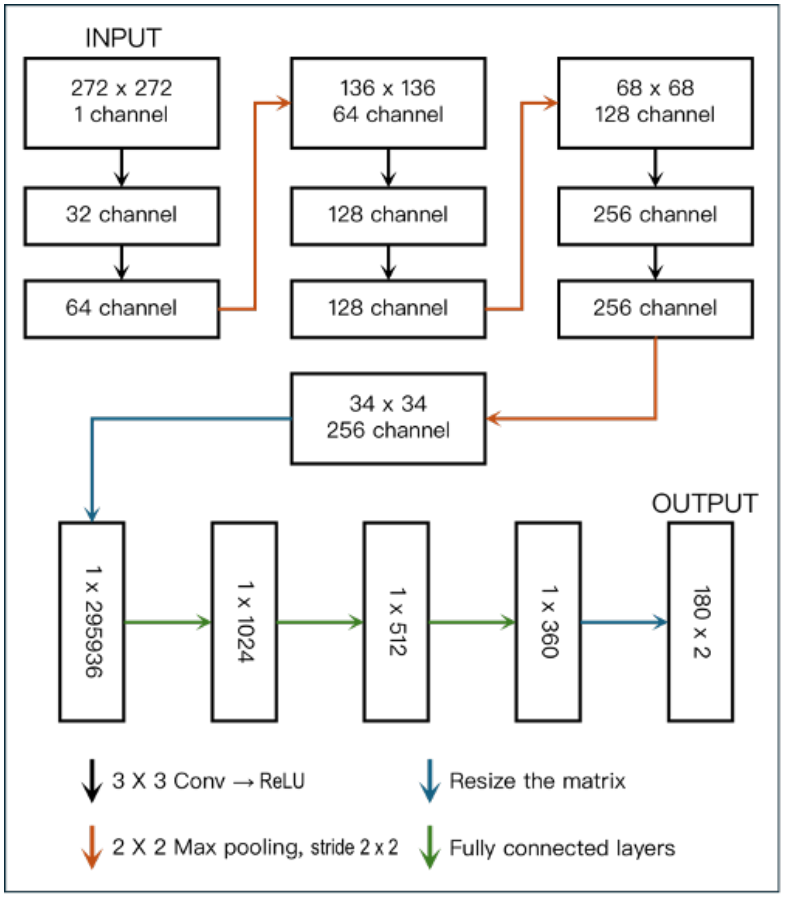


Figure 3.9: FCN Architecture

The more complex the model the more the model can find connections from different kinds of features extracted in the hidden layer, including the outlier and noise so to produce a model that is more robust to outlier and input noise, the simpler model was used.

FCN architecture as seen in Figure 3.9 is the proposed neural network model that is used to generate the error-free contour data. This model will produce an articulator contour that is made from 180 format-point contour. We chose this format because this format comprises more than 50% of available contour data as detailed in Table 3.2. The input of the FCN is the video frame of MRI data that has been processed into higher resolution using an Enhanced Deep Residual Network (EDRN) [54] to get a more accurate edge of the articulator as can be seen in Figure 3.10.

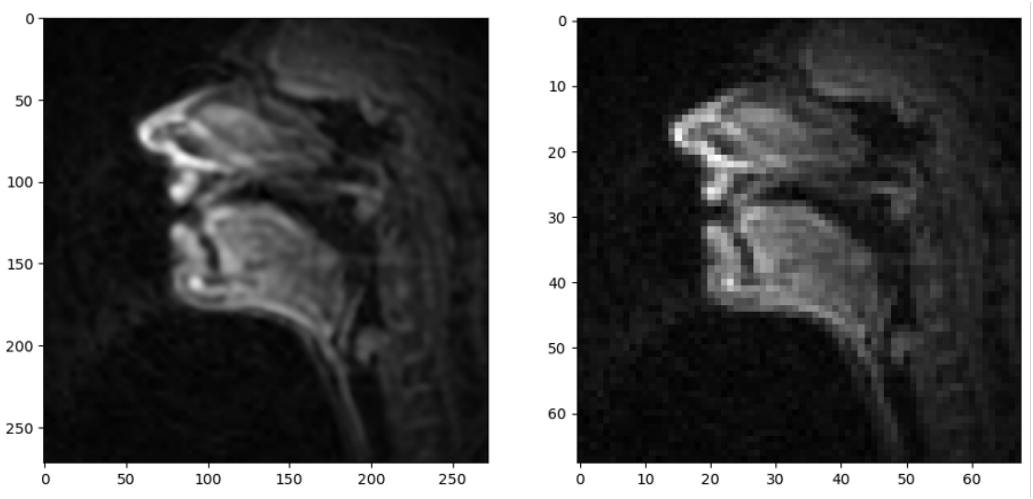


Figure 3.10: Different between super-resolution (left) and original frame (right)

Convolutional Layer

The convolutional layer works by iterating several filters of equal sizes across the image to look for special patterns in the image, the filter will create a new grid to highlight the patterns that are found by the kernel. The training process will highlight the pattern that contributes to the accuracy of output and decrease the pattern that is not important. Then the layer will go through the activation layer ReLU which helps the network learn non-linear relationships between image features, enhancing its robustness in identifying diverse patterns. Additionally, it aids in mitigating vanishing gradient problems, ensuring more effective training of the network.

Pooling Layer

Pooling layer was added to filter the feature form convoluted matrix. It also reduce the dimension of convolutional layer. Max pooling is used to get the maximum value of the feature map.

Fully Connected Layer

The output of pooling layer then flattened into 1D. ReLU activations functions are applied. Then, to generate probability values for each of the possible output labels a softmax layer was added. The predicted label is the output with highest probability score.

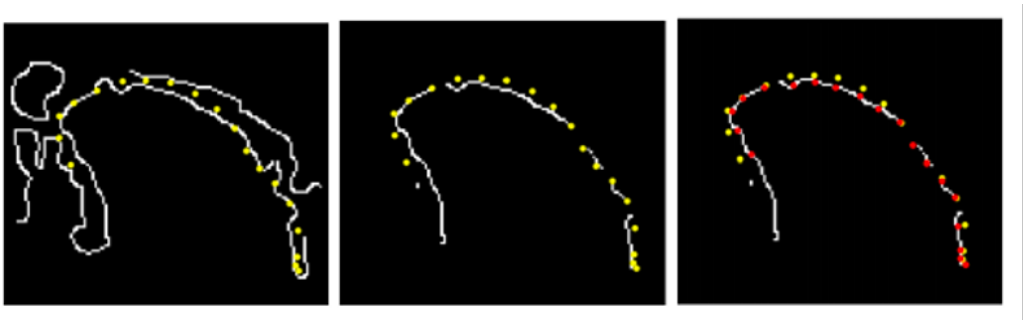


Figure 3.11: Yellow dots are the output of FCN, process left to right; (left) White lines as edges from edge detection; (middle) Only retain the necessary edge lines; and (right) Projecting the yellow dots onto the line, represented as red dots.

3.4.3 Point-to-curve Projection

The output of the FCN model is the generalized output. As it is generalized, the model often ignores the rough edge in the dataset and produces smoother contours that are away from the actual edge of the articulator contour, as seen in the example Figure 3.11 (left). So to keep the contour on the edge of the articulator, contours are projected into the articulator's edge. Articulator edge is generated using the adaptive threshold Gaussian method [55]. After only retaining the articulator contour area edge, Figure 3.11 (middle), the contour points are then projected into the contour edge, Figure 3.11 (right).

3.5 Experimental Setup

3.5.1 Model Parameter

PyTorch 2.0.1 [56] was used to implement the model, and training was conducted on an NVIDIA GeForce RTX 3090 graphics card. MRI image was upsampled into super-resolution of 272×272 pixels using Single-scale SR Network (EDSR) [57], this was done to increase the clarity of edge. While the output comprised the 360 sequence data that represent x and y coordinates of 180 landmarks. The Adam optimizer [58] was utilized with hyperparameters $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\epsilon = 1e - 4$ and 100 epochs.

3.5.2 Subjects Evaluation

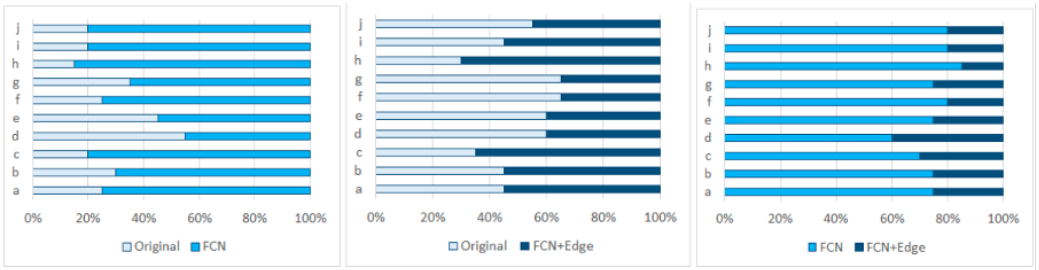
Subjective evaluation was conducted with 20 participants, to compensate the lack of gold-level ground truth of the contour data. The method of

the subjective is an A/B preference test, in which the participant needs to choose between 2 displayed contours on which contour better represents the vocal tract contour. There are 3 pairs of comparison group, original contour dataset (original), output data of the FCN model (FCN), and the output of the point-to-edge projection (FCN+Edge). Those groups are compared to each other which results in 3 comparison groups: Original vs FCN, Original vs FCN+Edge, and FCN vs FCN+Edge with each group has 20 questions each.

The participant did not have the knowledge of 3 kinds of labels, to avoid bias and complexity. Randomly selected label pairs were set to be shown to the participants to be compared (A and B). On the image, articulator contours are grouped into nine areas, as seen in Figure 3.7. Participants are then asked to compare each area from two given images, in which contour gave the less error or more accurate contour representation. This experiment was conducted using google form[1].

3.6 Experimental Result

3.6.1 Subjective Evaluation Result



(a) Original vs. FCN (b) Original vs. FCN+Edge (c) FCN vs. FCN+Edge

Figure 3.12: Comparison of preference between 3 pairs of groups, evaluating nine areas as well as the overall contour.

The result of AB preference tests between two contours is detailed in Figure 3.12, with the alphabet character a-j representing the contour of the upper lip, bottom lip, hard palate, edge of the tongue, middle tongue, back of the tongue, epiglottis, uvular, pharyngeal wall, and the overall landmark-based vocal tract. Figure 3.12 shows that the labels from only the smoothing

[1]A survey administration software included as part of the free, web-based Google Docs Editors suite offered by Google.

step for overall contour data are chosen 80% of the time, with the improvement of almost all contour areas compared to the original data, except on the edge of the tongue area, which is on this area, original data was chosen 55% of the time. For the preference test between FCN-only labels and FCN+Edge labels, also FCN-only labels are preferred in all contour areas. As for FCN+Edge labels, it only gives an improvement in the hard palate and uvular areas as it was chosen 65% and 70% respectively. For other areas, especially on the tongue areas original dataset was preferred compared to FCN+Edge.

3.6.2 Discussion

The result of the subjective evaluation showed that the best labels out of the step proposed in refining the original dataset are the labels from FCN-only labels. The subjective evaluation also showed that the proposed refinement method gave very high improvement (>60%) in 7 areas: upper lip (yellow), lower lip (purple), hard palate (pink), back of tongue (orange), epiglottis (blue), uvular (grey), pharyngeal wall (brown), but it only gave slight improvement for edge and middle of tongue. This is the expected result as the smoothing process using FCN eliminates the occasional errors and generalizes the contours. The effect of generalization made the output in the most moving part (edge and middle of the tongue) to be less precise than the actual edge of the contour. The label resulting from the edge projection step did not give any improvement compared to the original, this happened because the method that was used to generate the original dataset and the edge projection step relied on the edge detection from the image frames, and in many cases, the MRI videos have blurry images, which resulted in both methods to produce labeling error.

The common errors and inaccuracies that are found on the group labels can be seen in Figure 3.13. For the original dataset, the errors happened on the uvular area (grey) and the epiglottis area (blue), and also commonly labeling errors happened when 2 contour areas were close to each other or even touching, as it resulted in the edge detection method confused in determining the edge, this error example can be seen in (a) left circle between tongue (red and green) and hard palate (pink). For the FCN-only labels, the inaccuracies happened because of the smoothing process, for the most moving contour edge and middle tongue (red and green), the contours have inaccuracies and have less precision to the actual edge of contour. For the FCN+Edge labels, because it reproduces a similar error from the original dataset both of them relied on the edge detection method.

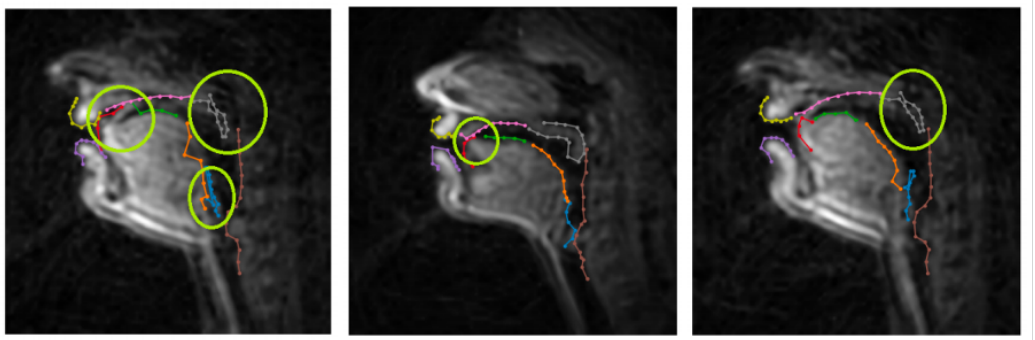


Figure 3.13: Errors and inaccuracies of the landmark contour labels identified from each group.

3.7 Summary

The proposed method refines landmark-based vocal tract contour labels through a series of steps, including outlier removal, FCN-based smoothing, and landmark point-to-edge curve projection. The quality of the proposed approach was compared to original dataset by subjective evaluation although no established ground truth labels exist, the quality of the newly refined labels was assessed subjectively by comparing various contour areas to the original data labels. The results indicate that the FCN-only labels significantly outperform both the original labels and the FCN+Edge labels. The uvular region, which was particularly prone to errors in the original dataset, showed notable improvements. Overall, the refined labels, incorporating outlier removal and FCN-based smoothing, greatly improve accuracy and reliability, providing enhanced vocal tract label data that is publicly available[2]. These FCN-smoothed labels will be used for training articulator movement models.

[2]The refined rtMRI landmark-based vocal tract contour label data proposed in this study is available at `https://github.com/ha3ci-lab/USC-TIMIT_rtMRI_Landmarks` and can serve as auxiliary information for the existing USC-TIMIT dataset.

Chapter 4

Proposed Articulator Movement Generator

4.1 Overview

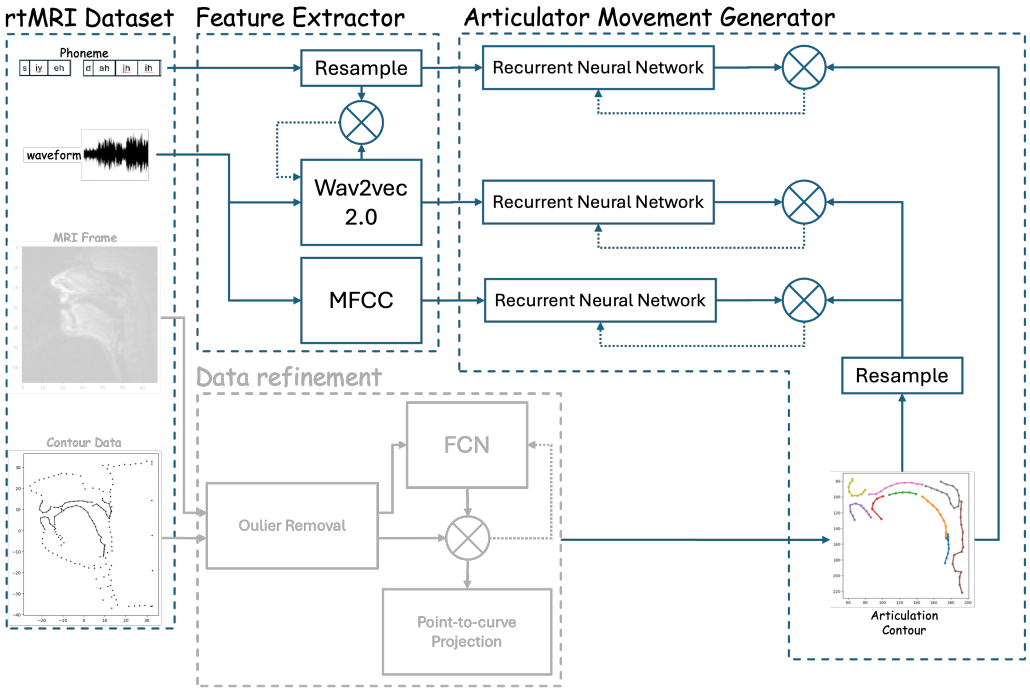
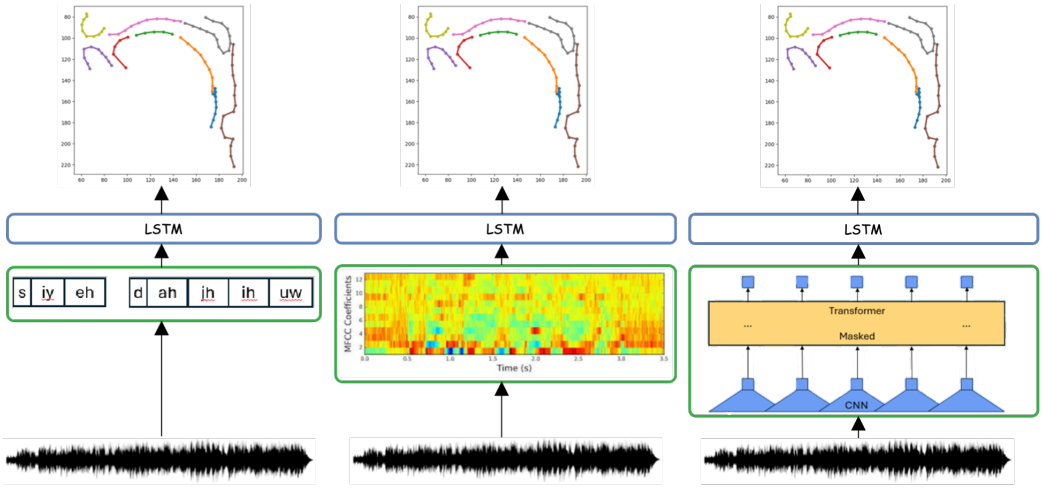


Figure 4.1: Articulation Movement Generator Section of Proposed Method

This study proposes a model that will generate the articulator movement animation from speech input, and then the output will be used in pronunci-

ation training. The proposed method for an articulator movement generator consists of 2 steps: features encoding and movement generator model training. The overall model structure is shown in Figure 4.1, and the 3 proposed feature extraction methods are shown individually in Figure 4.2.



(a) (b) (c)

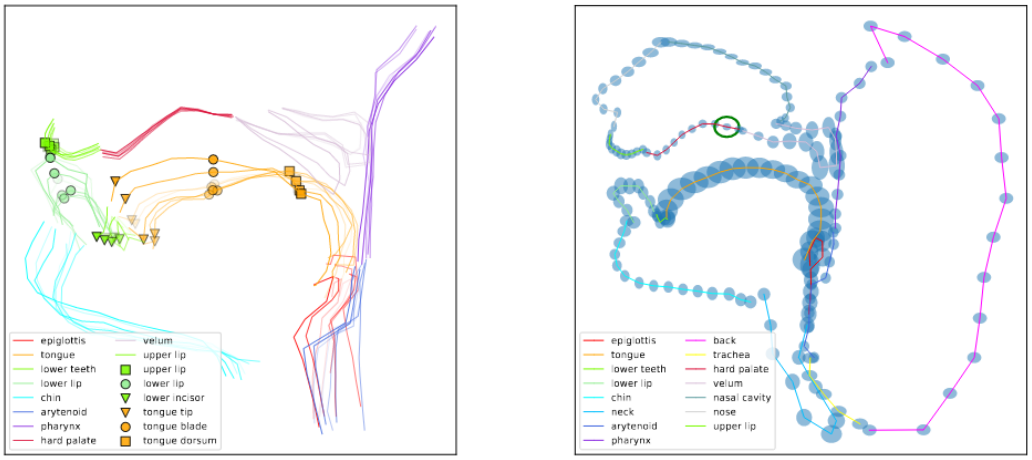
Figure 4.2: 3 Model with different Feature Extraction Methods

Feature encoding step is to extract features from the speech signal. The second step is the articulator movement generator model training. To generate the feature of speech, 3 different feature extractors will be evaluated. The first baseline feature is the human-annotated phoneme label. Phoneme definition is "any set of similar phones (speech sounds) that is perceptually regarded by the speakers of a language as a single distinct unit, a single basic sound, which helps distinguish one word from another" [59]. The second baseline feature is MFCC feature, a widely used speech feature that represents frequency bands that are equally spaced in the mel scale which is closer to the human auditory response. The third is the the proposed feature extraction method using one of the state-of-the-art ASR models, wav2vec 2.0 [60], that is capable of producing its context representation of speech sound. For the wa2vec2 feature extraction method additional fine-tuning stage will be conducted.

The second step after feature extraction is the contour movement generation using the recurrent neural network that could learn the temporal information in the data. The movement of the articulator's contour is affected by the previous position of the contour, which makes the recurrent network model necessary to retain the temporal information of the movement of the articulator's contours.

4.2 Related Works

4.2.1 Articulatory to Speech Conversion



(a) (b)

Figure 4.3: (a) Extracted MRI features. From EMA features the labeled points are estimated. (b) The extracted contour with a standard deviation denoted as the circle size of each point. [9]

Wu et al. [9] proposed an articulatory-to-speech model using the deep neural network (DNN). The data that was utilized is rtMRI data of native American English speaker woman, with the length of speech 17 minutes. Articulation contour was detected using a semiautomatic method [47] to track the contours of vocal tract air tissue boundaries from each frame (Figure 4.3(b)) and segmented the contours into anatomical components, as shown in Figure 4.3(a). Data comprised of xy-coordinate that when flattened for training become 230 dimensional-vector. The unnecessary contour were pruned.

Model

The baseline method was CNN-BiLSTM (CBL) [61]. The input of video frame passed into CNN and max-pooling layers, extracting the feature as BiLSTM input to produce mel-spectrogram output. Using waveform signal was generated using neural vocoder, HiFi-CAR [62], derivative of the HiFi-GAN convolutional network [63].

Denoising

The available dataset has significant reverberation and noise, by using off-the-shelf Adobe Podcast toolkit3[1] the quality of the speech recordings was enhanced.

Synthesis Quality

This research evaluates the synthesis quality and speech intelligibility using mel-cepstral distortions (MCD) [64] between ground truths and synthesized samples and character error rate (CER). Using Whisper [65], texts were generated from the synthesized speech for all test set utterances. The result was that the model outperformed both baselines. This shows that the task of generating sound from the articulator movement with mel ceptral distortion (MCD) 6.64 ± 0.64, and character error rate (CER) $69.2\% \pm 28.1\%$ made the inverse task seems probable.

4.2.2 Phoneme-to-rtMRI Video Generation

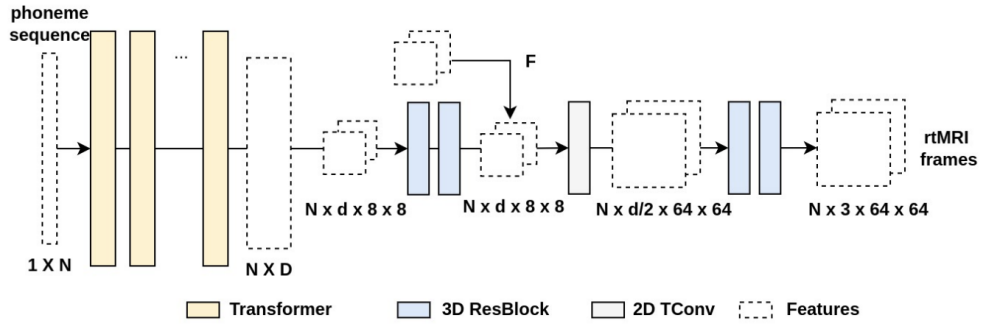


Figure 4.4: Model to generate rtMRI frames from phoneme sequence input. [10]

Udupa and Ghosh [10] proposed rtMRI video generation with the input of phonemes sequence and output of video frame representing rtMRI data. Using transformer architecture to encode the input phoneme sequence, convolutional neural networks (CNN) decoder to output rtMRI video frames. Upsampling output from the tranformers encoder using the CNN decoder maintaining the temporal context learned from the transformer encoder.

This research results in a model that is capable of producing the rtmRI videos with ABX test preference in range between 40-60% and mean opin-

[1]https://podcast.adobe.com/enhance

ion score between 3.4 - 4.16. This result shows that with the input of the phoneme, the encoder-decoder model is capable of generating rtMRI frames, and with more dimensional data that contain more information such as speech, it will improve the model with a similar structure.

4.2.3 Real-time Articulatory Visual Feedback with EMA

Suemitsu et al. [11] proposed a real-time articulatory movement feedback for pronunciation learning. The system was using EMA instrument as real-time input of the articulator movement as seen in the 4.5a. Sensors were placed in several places using EMA setup 4.5b, then visual output was displayed 4.5c. Sensors were placed on the tongue tip (TT), blade (TB), dorsum (TD), lower incisors (LI), upper lip (UL), and lower lip (LL). To compensate for head movement there are reference sensors on the upper incisors, nasion, and mastoid processes.

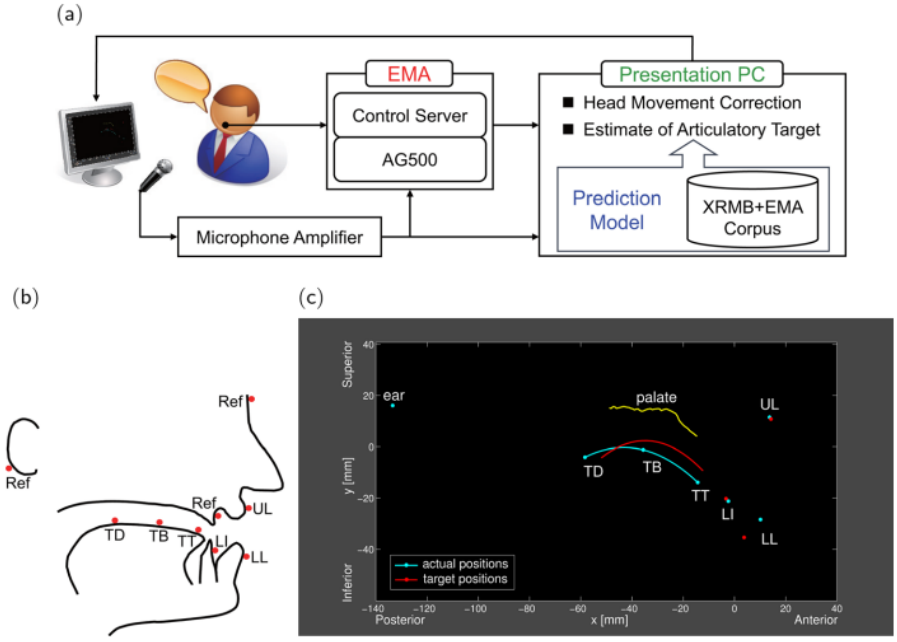


Figure 4.5: (a) System flow. (b) EMA setup. (c) Real-time visual feedback with visual display. [11]

1. V condition: visual feedback of tongue position with no acoustic cue
2. A condition: acoustic cue with no visual feedback
3. VA condition: visual feedback presented with acoustic cue

Table 4.1: Comparing the Japanese pronunciation with American English speakers for the pronunciation of vowel /æ/ in 3 conditions

Condition	Visual feedback	Acoutic cue
V	✓	×
A	×	✓
VA	✓	✓

The result of the experiment shows that learners in V and VA conditions produce /æ/ pronunciation that is closer to the native center position of /æ/, and this does not happen on A condition. This result shows that using the real-time visualization cue or feedback as a short-time learning session improves the pronunciation of learners.

Even though this study shows the improvement in pronunciation training using real-time visual cues, this study only investigates a specific phoneme and is limited to an EMA sensor that could only cover several areas of mouth articulation, the wider range of phoneme coverage and more information in articulator movement could improve the pronunciation training system.

4.3 Proposed Speech to Articulator Movement Generator

4.3.1 Overview of Self-Supervised Learning

Getting high accuracy in neural network models, large quantities of data has a very significant impact. But in many cases, the unlabeled data is easier to find and more in quantity compared to labeled data. Training high-accuracy model speech recognition systems normally requires a lot (thousands of hours) of labeled speech to reach acceptable performance. But, this approach is different than how humans learn languages, as an example of an infant that learns from its surroundings. This process will lead the infant to learn a good representation of speech by himself, or self-supervised. Self-supervised learning become a paradigm for model training that does not rely on labeled data but learns general data representations from unlabeled examples and then the model is fine-tuned on labeled data.

4.3.2 MFCC versus Wav2vec 2.0 Feature

Mel Frequency Cepstral Coefficients (MFCCs) and Wav2Vec are both techniques used in the processing and analysis of audio signals, particularly in the fields of speech recognition and audio analysis. However, they operate based on different principles and are suited to different tasks within the domain of audio processing. Comparison of the two methods:

MFCC

MFCC is the coefficient from a collective meal-frequency cepstrum. Cepstrum is the result of the inverse Fourier transform of the logarithm of the estimated signal spectrum. This tool is to investigate the periodic structure of in-frequency spectra.

Origin and Usage: Developed in the 1980s, MFCCs are one of the most popular feature extraction techniques used in automatic speech and speaker recognition. They effectively represent the vowel sounds which are important for understanding human speech. MFCC frequency bands are equally spaced in the mel scale which is closer to human auditory response than the linear-spaced frequency band spectrum.

There are 3 steps for MFCC feature extraction:

1. Speech signal is framed and windowed into 20-40 ms per frame, then Fourier Transform is applied to get the spectrum of the signal. The spectrum is squared to get the power spectrum.

2. Then we apply a triangular filter on a Mel-scale to the power spectrum to extract frequency bands. The logarithm of the mel spectrogram is taken to convert the spectrum into a log scale, which better approximates human hearing.

3. To convert the mel-spectrogram into a cepstral domain, Discrete Cosine Transform (DCT) it converts the sequence of data points into a sum of cosine functions oscillating at different frequencies.

Characteristics: MFCCs capture timbral/textural aspects of the sound, making them powerful for speaker identification and reducing the signal to a form that is less sensitive to the exact shape of the vocal tract.

Wav2vec 2.0

One of the state-of-art self-supervised learning framework for raw audio data is Wav2vec 2.0. This model was trained by masking the latent speech representation from multi-layer convolutional neural network. This representation

then passed into a transformer network to output contextualized representations. Wav2vec 2.0 was trained via a contrastive task between those 2 representation.

Origin and Usage: The initial model, Wav2Vec [66], was introduced in 2019. Wav2Vec 2.0 [60] is the modified model with a bigger parameters. This framework is used for speech recognition tasks, showing state-of-the-art performance by learning robust representations of speech from unlabeled data.

There are 2 steps in using this framework.

- Unsupervised Learning: The model learns directly from the raw audio waveform in an unsupervised manner initially. It uses a contrastive loss to align the latent speech representations with their contextualized representations.

- Fine-Tuning: After pre-training, the model is fine-tuned on a smaller amount of labeled data for specific tasks like speech recognition.

Characteristics: Wav2Vec2 models has the ability to effectively learn complex patterns in speech data without requiring manual feature engineering like MFCC. They leverage recent advances in neural networks to directly model the raw audio waveform, learning features automatically that are relevant for the task.

Comparison

- Performance: Wav2Vec generally offers superior performance for speech recognition tasks compared to traditional methods using MFCCs, especially in noisy environments or where the audio has characteristics not well covered by the training data for MFCC-based systems.

- Complexity and Resource Requirements: Wav2Vec models are significantly more complex and require considerably more computational resources both for training and inference.

- Flexibility and Adaptability: While MFCCs provide a fixed type of feature, Wav2Vec models are adaptable and can potentially learn any feature that is relevant to the task during the unsupervised training phase.

In summary, while MFCC remains a robust choice for many traditional audio processing tasks due to its simplicity and effectiveness, Wav2Vec represents a modern approach leveraging the latest in deep learning technology to

provide powerful, adaptable models especially suited to more complex speech recognition challenges.

4.3.3 Articulator Movement Generator utilizing Wav2vec 2.0

The proposed system uses using wav2vec2 system as the speech feature extractor, utilizing the self-supervised learning (SSL) capability of the wav2vec2 model to pre-train on large volumes of unlabeled audio data, resulting in the feature extraction method that outputs comprehensive speech representations. In this study, the capability of wav2vec2 in extracting the comprehensive speech representation is used to extract the speech feature. The model diagram can be seen in the Figure 4.1

The proposed Articulator Movement Generator model consisted of a Recurrent Neural Network that could learn the temporal information of sequence data. In this study, 2 stages will be used. For the first stage, the LSTM model will be used to compare 3 speech feature extraction methods. The 3 features will be used as input for LSTM model, with the output is a contour label that was extracted in chapter 3. Then, the model articulation movement generation capability will be evaluated. The second stage, using more complex model, Bi-LSTM to increase the accuracy of the model, proposed wav2vec2 features will be used to improve the accuracy of the articulator movement generator model.

4.4 Experimental Setup

The wav2vec2 fine-tuning and LSTM model training was implemented using PyTorch 2.0.1 [56], with the same machine as dataset refinement. The input is one of the 3 audio features (phoneme, MFCC, and wav2vec2), with the output in the same format as dataset refinement. Optimizer hyperparameter's different was only in the ϵ with $5e-5$. The network was trained for 500 epochs for every features and the best model that does not overfit will be chosen.

4.4.1 Dataset

2 kinds of data are used, speech data and articulation contour data. Speech data that was recorded simultaneously when articulator movement was scanned contained the noise that was the result of the data acquisition setup. The denoising method was applied to the dataset but it resulted in the removal

Table 4.2: Phoneme ground truth data resampled into the contour frame rate.

Phoneme	Train Data	Test Data	Phoneme	Train Data	Test Data	Phoneme	Train Data	Test Data
/uw/	3046	514	/s/	7232	1338	/dh/	1695	269
/aw/	1884	149	/z/	6029	937	/w/	1478	371
/ow/	3119	681	/l/	7508	1484	/ao/	3822	615
/b/	1791	209	/er/	6717	950	/uh/	823	146
/p/	2227	290	/y/	1032	217	/k/	3510	597
/m/	3086	450	/ay/	4709	952	/g/	1295	115
/f/	1618	201	/iy/	7398	1124	/ga/	3871	637
/v/	1335	236	/ey/	4978	644	/hh/	1127	264
/sp/	1335	90	/oy/	919	138	/ah/	10753	1710
/t/	6682	910	/ch/	1103	105	/ih/	9844	1467
/d/	3631	439	/zh/	247	31	/eh/	3911	729
/n/	7023	1128	/jh/	1554	143	/aa/	4464	663
/r/	6695	1304	/th/	480	62	/ae/	6390	1044
/s/	7232	1338	/sh/	1876	290	/ng/	1687	193
			/dh/	1695	269			

of several phoneme data including /hh/, /ah/, /ih/, /eh/, and /uh/ data. Because of this instead of denoising the speech data the feature extractor will be fine-tuned to adapt to the noise in the speech data.

Speech data are categorized into 2 groups, voice detected group and the silent group, the model then will be trained only using the voice-detected group and ignoring the silent group. This process is to make sure the model does not learn the silent stage when contour movement is unpredictable and does not have any pattern. The process of dividing the data into 2 groups was achieved using the labeled phoneme data ground truth that was annotated by humans. As shown in Figure 4.7, human-annotated phonemes contain the start and the end timeline of the phonemes, including the silent state.

The articulation contour data was derived from the refinement method explained in Chapter 3. The label from the output of the FCN-only model becomes the pair data with the speech data for the articulator movement generator training phase.

Speech data was down-sampled into 16kHz from the original 20kHz to match the waveform specification of the wav2vec 2.0. This setting will be applied globally for MFCC and wav2vec2 feature extraction.

The contour label was in the format of 180-point data with xy-coordinates, to train the model, the contour label data was flattened, resulting in the sequence of 1D data with a length of 360.

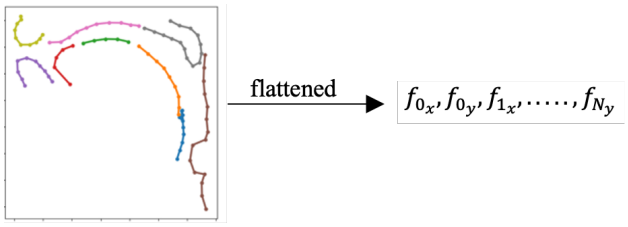


Figure 4.6: Contour label flattening

4.4.2 Baseline with Phoneme Sequence Feature

Phoneme sequences are the phoneme representation of audio input, which maps the section of the speech into a phoneme symbol. Each phoneme symbol is a single distinct unit, that represents a single basic sound, that distinguishes one word from another. The sounds that are perceived as phonemes differ between languages. An example case is in the phoneme sound of [n] in sin and [ŋ] sing are different phonemes in English but they constitute a single phoneme in Spanish, in which [pan] and [paŋ] are considered the same phoneme.

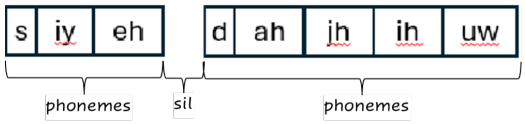


Figure 4.7: Phoneme feature example sequence

The phoneme used in the training steps was the human-annotated phoneme sequence. In this study a set of phonemes with the size of an N token is used as the input token for the phoneme sequence feature, additionally, the silent (sil) token is also added as a state when there is no speech data input. Padding token (PAD) and unknown (UNK) token are also used to normalize the size of the phoneme sequence for different lengths of sound and to handle the token that is not yet to be determined. The phoneme sequence is then used as the pair data with corresponding contour data.

The phoneme sequence with the shape of n occupies a certain timeline in the data point $(p^0, ..., p^n)$, with the sample rate following the speech sample rate. To be able to use the phoneme sequence for contour movement training, the phoneme sequence is then resampled into the same frame rate of the contour data, as shown in Figure 4.8. It will result in the repeated phoneme across its timeline $(p_0^0, ..., p_t^0, p_{t+1}^1, .. ., p_N^n)$, to provide the sequence of pair data for the phoneme and its corresponding contour movement of the size of $(f_0, ..., f_t, ..., f_N)$. The recurrent neural network is then trained with

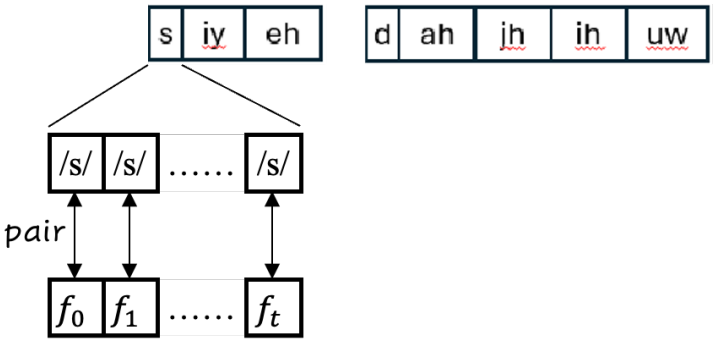


Figure 4.8: Resampling of phoneme sequence to match the size of contour data

the paired data, Figure 4.2 (a).

There are 41 defined phonemes as shown in Table 4.2, with additional silent (sil), padding (PAD), and unknown (UNK) tokens resulting in 44 tokens for training. 44 tokens then be converted into one-hot encoding for training. The contour data representing each phoneme can also be seen in Table 4.2.

4.4.3 Baseline with MFCC Feature

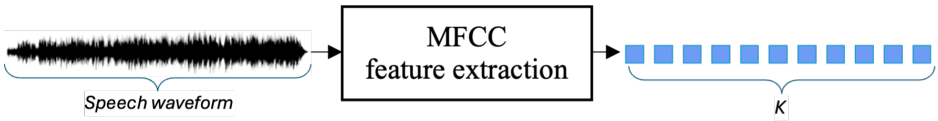


Figure 4.9: Feature extraction for MFCC

MFCC features are extracted from the speech data with a frame of 23ms, with frame-step automatically adjusted to produce the same feature sequence length output as wav2vec2, with the equation 4.1. The resulting feature will have the size of $80 \times K$. Because the frame rate of contour data is lower than the MFCC framing step. The result is the size of the MFCC feature will be longer than the length of contour data, t. To make the pairing data between MFCC feature and contour data, contour data will be resampled into the size of MFCC feature K. Then the features will be used for articulator movement generator model training, Figure 4.2 (b).

$$K = \sum_{i=0}^{7} f(i), \quad f(i) = \left\lfloor \frac{x-k}{s} \right\rfloor + 1 \tag{4.1}$$

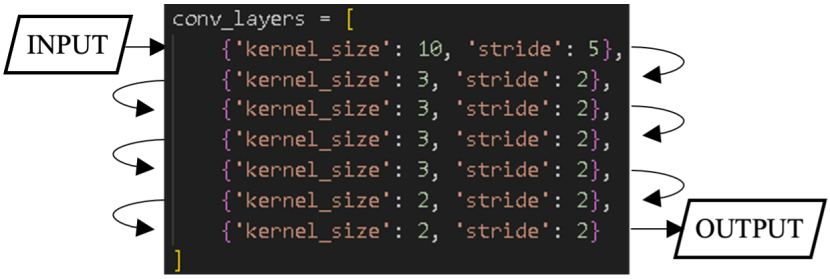


Figure 4.10: Convolution Layers Shape of Wav2vec 2.0

where $f(i)$ depends on x as follows:

$$x = \begin{cases} \text{input length} & \text{if } i = 0 \\ f(i-1) & \text{otherwise} \end{cases}$$

where,

- k is the kernel size
- s is the stride
- K is the feature length

4.4.4 Proposed System

Fine-tuning Strategy

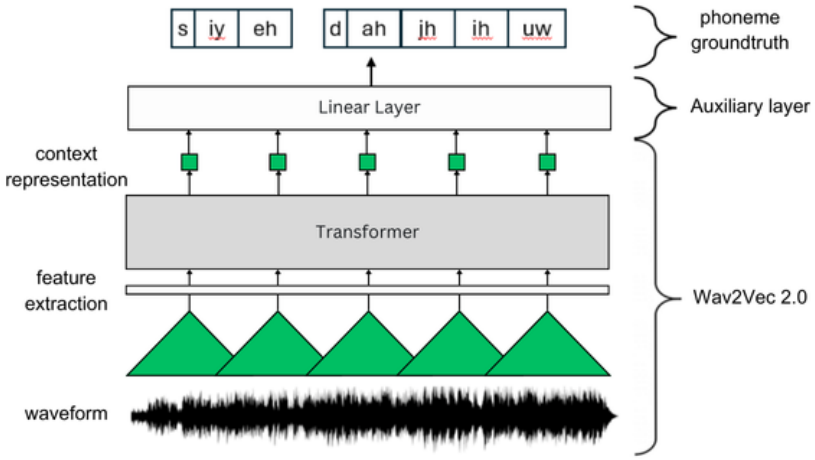


Figure 4.11: Checking the capability of wav2vec2 in generating phonemes

There are several steps for wav2vec2 model fine-tuning before the model is used to extract the speech feature for the articulator movement generator. rtMRI speech dataset was recorded simultaneously when the rtMRI data was scanned, resulting in the audio data that contained noises, so there is a need to check the capability of the model first before the model is used to extract the speech feature.

In the fine-tuning phase, the phoneme token output of wav2vec2 was decoded using a greedy search, equation 4.2, to find the token with the highest probability from the vocabulary, greedyCTC also be implemented to get the CTC phoneme sequence output.

$$\hat{y}_t = \arg\max_{y_t} P(y_t|x, \hat{y}_{1:t-1}) \tag{4.2}$$

where,

- x is input sequence
- y_t be the output token at time step t
- $P(y_t|x, y_{1:t-1})$ be the probability of y_t given the input sequence x and the previously generated tokens $y_{1:t-1}$.

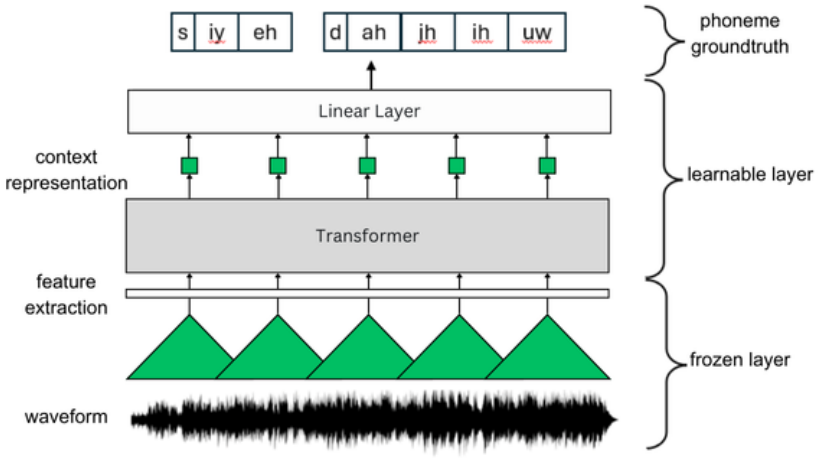


Figure 4.12: Fine-tuning wav2vec 2.0 with dataset data.

To check the capability of the wav2vec2 model in extracting the speech feature, the model output is evaluated in inference mode, Figure,4.11, using phoneme error rate (PER) metric to evaluate the performance of speech recognition systems at the phoneme level. The output phoneme of the model was compared with the phoneme ground truth. Then wac2vec2 model was

fine-tuned to adapt to the rtMRI phoneme datasets, Figure 4.12. Then, the model will be evaluated using PER to check the improvement. After that, the fine-tuned wav2vec2 model is used to generate audio features for articulator movement generator model training.

$$PER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{4.3}$$

where,

- S is the substitutions number
- D is the deletions number
- I is the insertions number
- C is the correct words number
- N is the phonemes in the ground-truth (N=S+D+C).

Training Strategy

Wav2vec2 features are extracted from the speech data. The resulting feature will have the size of $1024 \times K$. The result is the size of the wav2vec2 features will be longer than the length of contour data, t. To make the pairing data between the wav2vec2 feature and contour data, contour data will be resampled into the size of the wav2vec2 feature K. Then the features will be used for articulator movement generator model training, Figure 4.2 (c).

For the articulator movement generator training phase, 3 different features were trained using the same LSTM parameter to evaluate the performance of each feature in generating the articulation contour movement. The LSTM model will have 512 hidden states, with 1 LSTM layer. The input of LSTM will follow the feature size of each feature, 44 for phoneme features, 80 for MFCC features, and 1024 for wav2vec 2.0 features. The model is trained with Adam optimizer [58], and learning rate 5e-5. The model was trained with mean square error (MSE) loss to get the maximum likelihood estimation of the model. The contour label was in the format of 180-point data with xy-coordinates, to train the model, the contour label data was flattened, resulting in the sequence of 1D data with a length of 360. The contour data is normalized by dividing the coordinate by 25, as all the articulator contours are in the range of -25 until +25 in both of x and y coordinates.

Then, the best model the will be trained using more complex LSTM (bidirectional LSTM, with 2048 hidden state, and 3 layers) to enhance the capability of model in generating the articulator movement.

4.4.5 Objective Evaluation Setup

To objectively evaluate the capability of the model in producing the movement, we employ Spearman's rank correlation coefficient [67], to measure the correlation between 2 variables in monotonic function.

The formula for Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.4}$$

where:

- d_i is the difference between the ranks of corresponding values x_i and y_i.
- n is the number of observations.

4.5 Experimental Result

4.5.1 Wav2vec 2.0 Fine-tuning Task

Before the fine-tuning stage of wav2vec feature extraction, using the pre-trained model, the model was evaluated to have a PER of more than 60% using resampled phoneme sequence into contour length. On the training data, the PER of the model was 62.82% with CTC 24.41 and for the test data, the PER was 63.94% with CTC 24.87. This results in the inability of the model to extract the feature of the rtMRI speech data as more than 50% of the predicted output was wrong.

After the wav2vec 2.0 model's capability was evaluated, the pre-training phase as described in section 4.4.4 was conducted to adapt the wav2vec model to extract the audio features of rtMRI data. After fine-tuning, the PER of the model becomes 31.96% with CTC 0.11 for train data and 37.21% with CTC 0.43 for test data.

4.5.2 Feature Extractor Comparison

In this research to compare the different accuracies of model output we compare 23 features, from 4 categories, Lip Protrusion, Aperture, and position detail can be seen in Figure 4.13. For Lip Protrusion (LP), there are protrusions of the upper lip (Up) and protrusion of the lower lip (Down). For Aperture, there are aperture between 2 lips (L), aperture of tongue tip with the hard palate (TT), aperture of antero-dorsal with the palate (TAD), aperture of dorsal with the palate (TD), aperture of postero-dorsal with the soft

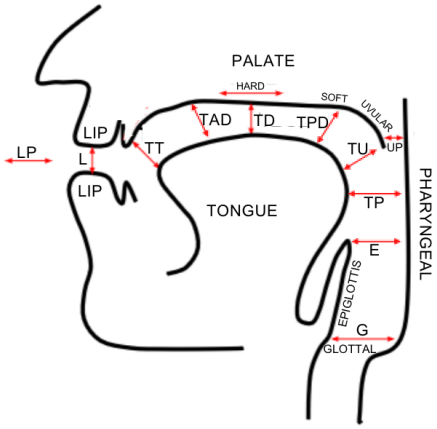


Figure 4.13: Extracted articulator features

palate (TPD), aperture of the tongue with the uvular (TU), aperture of uvular with the pharyngeal (UP), aperture of radical with the pharyngeal (TP), aperture of epiglottis (E), and aperture of the glottis (G).

Training the LSTM model using 3 different features input results in the model that can generate a movement with the spearman's correlation defined in the Table 4.3, the correlation between the generated contour movement and ground truth data. With the highest correlation from training using wa2vec2 features, except for glottal (G) aperture. From the 3 speech audio features that were used to train the LSTM model, the best output was the proposed model, wav2vec2 feature extraction. The model that was trained using this feature was capable of generating Then this feature was used to train the Bi-LSTM.

4.5.3 Result Analysis and Discussion

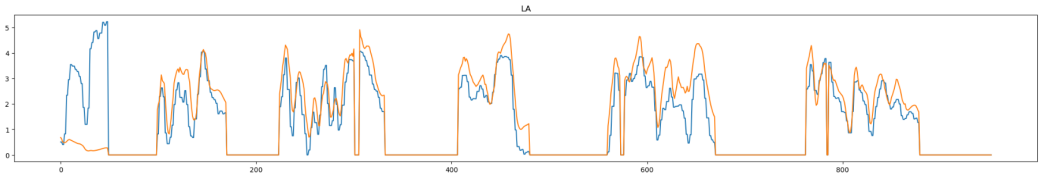


Figure 4.14: Example of lips aperture (LA) correlation in a sentence, The blue line is the ground truth of LA movement, and the orange is the predicted LA movement.

The result in the Table 4.3 shows the capability of the LSTM model in generating articulator movement with 3 different kind of speech features.

Table 4.3: Spearman's correlation between the generated movement and ground truth

Features		Wav2vec	MFCC	Phoneme
Lip Protrusion	**Up**	77.02%	72.54%	6.85%
	Down	64.91%	43.22%	47.72%
Aperture	**L**	67.75%	46.71%	43.89%
	TT	65.92%	43.72%	48.27%
	TAD	68.75%	44.48%	47.22%
	TD	61.81%	41.78%	47.01%
	TPD	64.05%	54.13%	55.09%
	TU	70.43%	48.92%	54.45%
	UP	84.61%	66.25%	57.89%
	TP	68.46%	56.46%	63.72%
	E	40.06%	35.04%	37.57%
	G	35.23%	50.53%	50.08%

Using wav2vec feature the model was able to produce the average correlation 64.08% compared to MFCC feature with 50.32% and phoneme feature 46.65%. This shows that for MFCC and phoneme features the model incapable to generate any movement as most of the time it just produces the average of the movement data resulting in near 50% correlation result.

The movement statistics over time for wav2vec2, as depicted in Figure 4.14, demonstrates that the LSTM model is capable of following the general trend of tongue movement. Employing a more complex LSTM model with 12× more parameters has improved the model's performance, achieving an average correlation of 67.76%. The most significant improvement was observed in the glottal (G) feature, with a substantial increase of 15.43%, reaching a 50% correlation. This correlation percentage indicates that the model can effectively generate movements for lip protrusion, lip aperture (L), tongue tip (TT), anterodorsal (TAD), dorsal (TD), posterodorsal (TPD), radical to uvular (TU), uvular to pharyngeal (UP), and radical to pharyngeal (TP). However, it still struggles to simulate movements for the epiglottis and glottis, with results from both LSTM and Bi-LSTM models falling below or near 50%—the average correlation level. This limitation may be due to the blurriness of these areas in MRI images, leading to potential errors in

Table 4.4: Spearman's correlation between the generated movement and ground truth

Features		LSTM	Bi-LSTM
Lip Protrusion	**Up**	77.02%	74.45%
	Down	64.91%	65.43%
Aperture	**L**	67.75%	68.97%
	TT	65.92%	73.44%
	TAD	68.75%	71.71%
	TD	61.81%	69.16%
	TPD	64.05%	70.54%
	TU	70.43%	74.53%
	UP	84.61%	82.33%
	TP	68.46%	70.00%
	E	40.06%	41.93%
	G	35.23%	50.66%

contour labeling.

4.6 Summary

The proposed model with wav2vec2 speech feature extraction needs 2 steps, The first is a fine-tuning stage, to adapt the feature extractor model (wav2vec2) into the rtMRI speech dataset, because the speech data contains noises resulting not fine-tuned model having very high PER. The second step is articulator movement generator training using the extracted feature. Compared to the 2 baseline methods, which failed to learn to generate the articulator movement, the proposed model is capable of generating a general trend of tongue movement, and using a model with a bigger parameter increases the accuracy of the model.

Chapter 5

Conclusion and Future Direction

5.1 Conclusion

This research was motivated by the universal principles underlying human sound production and its potential to aid pronunciation training. With the research aims to develop an articulator movement generation that could be used for a Computer-Assisted Pronunciation Training system. This research addresses several issues in current emotion speech conversion task technology:

- To detect articulatory movements corresponding to the recorded speech sound, including choosing the feature extraction methods, and dataset refinement.
- To assess the effectiveness of the articulator movement generator for the pronunciation detection task.

In this study, we built the articulator generation model for pronunciation detection to generate articulator movement with only speech input. This research produces a refinement method that was used to produce the necessary dataset for the pair data of speech sound and articulatory movements using rtMRI data.

From the 3 features that were used for model training, the feature that was capable of being used as training data for the articulator movement generator was the wav2vec2 feature after fine-tuning it with the speech dataset. While model training using the phoneme and MFCC features failed to even produce the mean of articulator movement. The model is also capable of following the general trait of the articulator movement, with a more complex model increasing the accuracy of the model.

The model's capability to detect pronunciation could be used for pronunciation training, as it is capable of generating the general trend of articulator

movement. This visualization feedback then can be used to enhance the learning ability of pronunciation of L2 learner, to detect errors in their pronunciation, or to compare themselves with the right pronunciation.

5.2 Future Direction

The future direction of this study will be:

- This study was conducted using only the English speaker dataset and using the English phoneme feature, this makes the produced model only capable of visualizing the model of English speech. By increasing the language that the model was trained it would be able to increase the capability of the model to understand multiple languages and achieve language agnostic model.
- The model was also trained using a simple LSTM model and its derivative, the Bi-LSTM model, which is still not state-of-the-art in sequence data learning. By adopting a more complex model and state-of-the-art model, the model's performance will also increase.
- Model has also only been trained using the rtMRI experimental setup audio that is very different than real-life situations. A new kind of dataset that is closer to real-world situations would help the model to adapt to real-world situations.
- The output resolution and frame rate of the model follows the rtMRI resolution and frame rate. For the purpose of language learning the there is a need of improvement for the resolution and frame rate that are effective in delivering information for language learning feedback.

List of Publications

M. R. Ridha and S. Sakti, "Refining rtMRI Landmark-Based Vocal Tract Contour Labels with FCN-Based Smoothing and Point-to-Curve Projection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 13796–13802. Accessed: Jul. 25, 2024. [Online]. Available: `https://aclanthology.org/2024.lrec-main.1204`

Bibliography

[1] Z. Cai, X. Qin, D. Cai, M. Li, X. Liu, and H. Zhong, "The DKU-JNU-EMA electromagnetic articulography database on mandarin and chinese dialects with tandem feature based acoustic-to-articulatory inversion," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 235–239, IEEE, 2018.

[2] J. Catford, *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press, 1977.

[3] Y.-T. Liu and W.-T. Tseng, "Optimal implementation setting for computerized visualization cues in assisting l2 intonation production," *System*, vol. 87, p. 102145, 2019.

[4] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," 2022.

[5] X. Qian, H. Meng, and F. Soong, "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1020–1028, 2016.

[6] F. Satria, H. Aditra, M. D. A. Wibowo, H. Luthfiansyah, M. Suryani, E. Paulus, and I. Suryana, "Efl learning media for early childhood through speech recognition application," in *Proceedings of the Third International Conference on Science in Information Technology*, pp. 568–572, 2017.

[7] Z. Raeesy, S. Rueda, J. K. Udupa, and J. Coleman, "Automatic segmentation of vocal tract mr images," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, (San Francisco, CA, USA), pp. 1328–1331, IEEE, Apr. 2013.

[8] M. Ruthven, M. E. Miquel, and A. P. King, "Deep-learning-based segmentation of the vocal tract and articulators in real-time mag-

netic resonance images of speech," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105814, Jan. 2021.

[9] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, "Deep speech synthesis from MRI-based articulatory representations," 7 2023.

[10] S. Udupa and P. K. Ghosh, "Real-time mri video synthesis from time aligned phonemes with sequence-to-sequence networks," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[11] A. Suemitsu, J. Dang, T. Ito, and M. Tiede, "A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning," *The Journal of the Acoustical Society of America*, vol. 138, pp. EL382–EL387, 10 2015.

[12] J. Fouz-Gonzalez, "Pronunciation instruction through twitter: the case of commonly mispronounced words," *Computer Assisted Language Learning*, vol. 30, no. 7, pp. 631–663, 2017.

[13] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *Journal of the Acoustical Society of America*, vol. 136, pp. 1307–1311, Sept. 2014.

[14] *Computer Assisted Pronunciation Training (CAPT): A Systematic Review of Studies from 2012 to 2022*, 2022.

[15] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (CAPT): Current issues and future directions," *RELC*, vol. 52, no. 1, pp. 189–205, 2021. Publisher: SAGE Publications Ltd.

[16] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in english," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, 2019.

[17] C. BERTORELLI, "Effectiveness of computer-assisted pronunciation training," *Journal of Tourism Studies*, p. 69, 2017.

[18] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5049–5052, 3 2012. ISSN: 2379-190X.

[19] R. Duan, T. Kawahara, M. Dantsujii, and J. Zhang, "Pronunciation error detection using DNN articulatory model based on multi-lingual and multi-task learning," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, IEEE, 2016.

[20] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

[21] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 2 2000.

[22] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic pronunciation error detection and guidance for foreign language learning," in *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia*, 11 1998.

[23] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The DISCO ASR-based CALL system: practicing l2 oral skills and beyond," in *Proceedings 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, May 2012*, pp. 2702–2707, 2012. Publisher: Istanbul, Turkey:[Sn].

[24] K. P. Truong, A. Neri, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in *Proc. InSTIL/ICALL 2004 Symposium on Computer Assisted Learning*, p. paper 032, 2004.

[25] P. Ladefoged, *The sounds of the world's languages*. Oxford, OX, UK ; Cambridge, Mass., USA : Blackwell Publishers, 1996.

[26] B. Lindblom and J. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," *The Journal of the Acoustical Society of America*, vol. 48, pp. 120–120, 07 1970.

[27] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, pp. 1070–1082, 1973.

[28] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling* (W. J. Hardcastle and A. Marchal, eds.), pp. 131–149, Dordrecht: Kluwer Academic Publishers, 1990.

[29] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips, and face, based on mri and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[30] B. J. Kröger and P. Birkholz, "A gesture-based concept for speech movement control in articulatory speech synthesis," in *Verbal and Nonverbal Communication Behaviours* (A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, eds.), pp. 174–189, Springer Berlin Heidelberg, 2007.

[31] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[32] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded gaussian mixture regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2246–2259, 2015.

[33] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[34] B. H. Story and I. R. Titze, "Parametrization of vocal tract area functions by empirical orthogonal modes," *Journal of Phonetics*, vol. 26, pp. 223–260, 1998.

[35] I. Stavness, P. Perrier, D. Demolin, and Y. Payan, "A biomechanical modeling study of the effects of the orbicularis oris muscle and jaw posture on lip posture," *Journal of Speech, Language, and Hearing Research*, vol. 56, pp. 878–890, 2013.

[36] P. Perrier, Y. Payan, M. A. Nazari, N. Hermant, P.-Y. Rohan, C. Lobos, and A. Bijar, "Speech biomechanics: What have we learned and modeled since joseph perkell's tongue model in 1974?," *Journal of the Acoustical Society of America*, vol. 139, no. 4, p. 2193, 2016.

[37] C. A. Fowler and E. L. Saltzman, "Coordination and coarticulation in speech production," *Language and Speech*, vol. 36, pp. 171–195, 1993.

[38] J. Wang, J. R. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7785–7789, 2013.

[39] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, p. S56, 8 2005.

[40] H. K. Vorperian, S. Wang, M. K. Chung, E. M. Schimek, R. B. Durtschi, R. D. Kent, A. J. Ziegert, and L. R. Gentry, "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," *The Journal of the Acoustical Society of America*, vol. 125, no. 3, pp. 1666–1678, 2009. Publisher: AIP Publishing.

[41] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992. Publisher: Acoustical Society of America.

[42] D. J. Olson, "Benefits of visual feedback on segmental production in the l2 classroom," *Language Learning and Technology*, vol. 18, 10 2014. Publisher: University of Hawaii National Foreign Language Resource Center.

[43] P. Wu, L.-W. Chen, C. J. Cho, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 6 2023. ISSN: 2379-190X.

[44] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[45] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *INTERSPEECH*, 2021.

[46] A. W. Black, *MOCHA-TIMIT*. Centre for Speech Technology Research, University of Edinburgh, Nov. 1999. Accessed: Jul. 25, 2024.

[47] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, 2008.

[48] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, Jan. 1988.

[49] M. Labrunie *et al.*, "Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning," *Speech Communication*, vol. 99, pp. 27–46, May 2018.

[50] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[51] A. Falcao, J. Udupa, S. Samarasekera, S. Sharma, B. Hirsch, and R. Lotufo, "User-steered image segmentation paradigms: Live wire and live lane," *Graphical Models and Image Processing*, vol. 60, no. 4, pp. 233–260, 1998.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), vol. 9351 of *Lecture Notes in Computer Science*, (Munich, Germany), pp. 234–241, Springer, 2015.

[53] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time mri sequences," *Computer Speech Language*, vol. 33, pp. 25–46, Sept. 2015.

[54] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," July 2017. Accessed: Oct. 21, 2023.

[55] S. Gaur, J. Vajpai, and S. Mehta, "Adaptive local thresholding for edge detection," *International Journal of Computer Applications*, vol. 2, p. 8887, Sept. 2014.

[56] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," Dec. 2019. Accessed: Oct. 21, 2023.

[57] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *arXiv preprint arXiv:1707.02921*, 2017.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 2017. Accessed: Oct. 21, 2023.

[59] Merriam-Webster, "Definition of phoneme." Accessed: Jul. 27, 2024.

[60] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2024. Accessed: Jul. 25, 2024.

[61] Y. Yu, A. H. Shandiz, and L. Tóth, "Reconstructing speech from real-time articulatory mri using neural vocoders," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, pp. 945–949, 2021.

[62] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep speech synthesis from articulatory representations," in *Proceedings of the International Speech Communication Association, INTERSPEECH*, pp. 779–783, 2022.

[63] J. Kong, J. Kim, and J. Bae, "Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[64] A. W. Black, "Cmu wilderness multilingual speech dataset," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5971–5975, 2019.

[65] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org, 2023.

[66] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (Graz, Austria), pp. 3465–3469, ISCA, 2019.

[67] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.