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.