Master's Thesis of Arts

A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis

자동발음평가-발음오류검출 통합 모델

August 2023

Graduate School of Humanities
Seoul National University
Linguistics Major

Hyungshin Ryu

A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis

Advising Professor, Dr. Minhwa Chung

Submitting a master's thesis of Linguistics

August 2023

Graduate School of Humanities
Seoul National University
Linguistics Major

Hyungshin Ryu

Confirming the master's thesis written by
Hyungshin Ryu
August 2023

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)

Abstract

Ryu, Hyungshin
Department of Linguistics
Graduate School
Seoul National University

Empirical studies report a strong correlation between pronunciation scores and mispronunciations in non-native speech assessments of human evaluators. However, the existing system of computer-assisted pronunciation training (CAPT) regards automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD) as independent and focuses on individual performance improvement. Motivated by the correlation between two tasks, this study proposes a novel architecture that jointly tackles APA and MDD with a multi-task learning scheme to benefit both tasks. Specifically, APA loss is examined between cross-entropy and root mean square error (RMSE) criteria, and MDD loss is fixed to Connectionist Temporal Classification (CTC) criteria. For the backbone acoustic model, self-supervised model is used with an auxiliary fine-tuning on phone recognition before multi-task learning to leverage extra knowledge transfer. Goodness-of-Pronunciation (GOP) measure is given as an additional input along with the acoustic model.

The joint model significantly outperformed single-task learning counterparts, with a mean of 0.041 PCC increase for APA task on four multi-aspect

scores and 0.003 F1 increase for MDD task on Speechocean762 dataset. For the joint model architecture, multi-task learning with RMSE and CTC criteria with raw Robust Wav2vec2.0 and GOP measure achieved the best performance. Analysis indicates that the joint model learned to distinguish scores with low distribution, and to better recognize mispronunciations as mispronunciations compared to single-task learning models.

Interestingly, the degree of the performance increase in each subtask for the joint model was proportional to the strength of the correlation between respective pronunciation score and mispronunciation labels, and the strength of the correlation between the model predictions also increased as the joint model achieved higher performances. The findings reveal that the joint model leveraged the linguistic correlation between pronunciation scores and mispronunciations to improve performances for APA and MDD tasks, and to show behaviors that follow the assessments of human experts.

Keyword: computer-assisted pronunciation training, multi-task learning, self-supervised learning, Goodness-of-Pronunciation, automatic pronunciation assessment, mispronunciation detection and diagnosis

Student Number : 2021-20387

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

The incorporation of speech technology into education has consistently grown and has brought meaningful results (Eskenazi 2009; Litman, Strik, and Lim 2018). The field of computer-assisted pronunciation training (CAPT) has similarly made rapid progress, with the spread of internet-based applications and the significant development of automatic speech recognition technology. The CAPT system serves as a powerful tool for non-native learners, as it provides customized feedback at a low cost. Minimized time and place constraints typical in traditional instructor-based learning bring another advantage to CAPT (Rogerson-Revell 2021).

The CAPT system generally consists of two major tasks, automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). APA task can be seen as a speech classification or regression task, which aims to provide pronunciation scores that are highly correlated with those of human evaluators (Gong et al. 2022; Naijo, Ito, and Nose 2021). The task models various types of scores, as the labeled scores reflect different aspects of scoring standards (phoneme, rhythm, intonation) as in ERJ dataset (Minematsu et al. 2004) or different granularities (phones, words, sentences) as in Speechocean762 (Zhang et al. 2021). MDD task on the other hand is a non-native phone recognition task. It aims to correctly classify and diagnose the recognized phones into correct pronunciations and mispronunciations, by comparing the recognized phones with the annotated phone transcriptions of human experts and the canonical phone sequences (Leung, Liu, and Meng 2019; Peng et al. 2021). The detected mispronunciations are classified into substitution, insertion, and deletion, and are further diagnosed into discrete

phones.

As APA and MDD both assess non-native (L2) speech, the two tasks inevitably share similar methodologies in acoustic models and acoustic features. Studies propose acoustic models that better reflect the general characteristics of non-native speech. Recently with the advance of deep learning, self-supervised learning (SSL) models are frequently adopted in MDD tasks after its first exploration in Xu et al. (2021) and Peng et al. (2021), followed by APA tasks in Chao et al. (2022) and Kim et al. (2022) for their robust acoustic representation. Acoustic features on the other hand capture a specific quality of speech. Goodness-of-Pronunciation (GOP) measure, a duration normalized posterior phone probability, is a representative acoustic feature proposed by Witt and Young (2000) to provide phone scores. The measure is used both in APA task (Sudhakara, Ramanathi, Yarra, and Ghosh 2019; Tong et al. 2014) and MDD task (Hu, Qian, and Soong 2015) for their high correlation with human experts.

Indeed, empirical studies report that there exists a distinct correlation between pronunciation scores and mispronunciations that are annotated by human evaluators for non-native speech assessments (Chen et al. 2016; Munro and Derwing 1995; O'Brien 2014; Yang and Chung 2017). Mispronunciations showed a strong correlation with not only overall assessment such as comprehensibility scores of L2 German (O'Brien 2014) and holistic scores of L2 Korean (Yang and Chung 2017), but also prosodic assessment such as fluency scores of L2 Mandarin (Chen et al. 2016), and accent scores of L2 English (Munro and Derwing 1995). This applied to both cases where mispronunciation annotators and score annotators were different (Chen et al. 2016; Munro and Derwing 1995; Yang and Chung 2017) or the same

(Yang and Chung 2017). This provides strong linguistic motivation to leverage the correlation between APA and MDD tasks to benefit each other.

However, current CAPT studies have treated the two tasks as independent and separate. One reason lies in the proposals focusing on improving the model performance on different benchmark datasets for the respective task. L2-ARCTIC (Zhao et al. 2018) is often used to test MDD performance, while Speechocean762 (Zhang et al. 2021) is used to measure APA performance. Other studies use Speechocean762 only as an MDD benchmark (Wadud, Alatiyyah, and Mridha 2023; Zhang et al. 2022). A few studies that mention both tasks still regard one as an auxiliary task to support the other (Lin et al. 2020; Lin and Wang 2023) or a task that can be separately achieved with a change in input or task-specific layers (Fan et al. 2023; Zheng et al. 2022). However, given the significant linguistic correlation, an integrated model of the two tasks is expected to improve performances on both tasks.

This work presents for the first time an integrated architecture that jointly trains pronunciation assessment task and mispronunciation detection and diagnosis task via a multi-task learning perspective, to leverage their correlation. Three architectures are experimented for the joint model. To further enhance the acoustic representation of the model, self-supervised learning model is used with additional phone recognition fine-tuning before multi-task learning. Four types of self-supervised learning model and three types of phonetically labeled datasets are experimented. This work contributes by verifying that the joint model shows distinct improvement on both APA and MDD tasks, compared to the respective single-task learning on Speechocean762 dataset. Analyses reveal that the performance improvements of the joint model with respect to single-task learning model are

proportional to the strength of the correlation between the labels, and that the correlation between model predictions also increases as the joint model achieves improved performances, which prove the importance of this study.

The remaining chapters are organized as follows: Chapter 2 provides an overview of the approaches in APA and MDD studies that are related to this work. Chapter 3 demonstrates the model architecture and experiment settings. Chapter 4 presents the experiment results of three different architectures for the joint model, and compares the results with respective APA and MDD model. Effects of leveraging different backbone models are also presented. The experiment results are analyzed for the model that had the best performance. Chapter 5 discusses how the joint model was able to leverage the correlation between pronunciation scores and mispronunciations to achieve performance improvements, using correlation analysis and experiment results of different multi-task learning loss weight. Chapter 6 summarizes the findings of the paper and their importance.

Chapter 2. Related work

Automatic pronunciation assessment task and mispronunciation assessment tasks share the main goal of assessing non-native speech using machine learning techniques. As the two tasks are primarily interested in building models that can exactly identify the differences in pronunciation, they inevitably share similar methodologies.

2.1. Acoustic models

Acoustic modeling is a major research subject in APA and MDD studies as it is directly related to extracting acoustic representations from utterances. The goal of acoustic modeling is to adopt robust models that can correctly reflect the characteristics of non-native speech. It takes advantage of the development in automatic speech recognition (ASR) system. The backbone acoustic models used in APA and MDD studies leverage the ASR architectures that are proven to show lower word error rates and robustness in other downstream speech tasks.

Traditional acoustic models include Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM)-based GMM-HMM. With the advance of deep neural networks (DNN), deep networks have shown to give significantly lower word error rates than the Gaussian networks (Hinton et al. 2012). Thus, the GMM module was substituted to the DNN module as to build DNN-HMM acoustic models. APA studies use these traditional acoustic models to obtain recognized non-native speech, which the machine learning or deep learning-based scoring module use as inputs to

output final scores (Lin and Wang 2021b; Metallinou and Cheng 2014). Similarly MDD studies use these acoustic model to get recognized phone predictions (Harrison et al. 2009; Tsubota, Dantsuji, and Kawahara 2004). However, GMM-HMM and DNN-HMM acoustic models in APA and MDD studies commonly have the burden of forced alignment for senone labels (Metallinou and Cheng 2014), pre-defined mispronunciation patterns (Harrison et al. 2009; Tsubota et al. 2004), and pronunciation dictionary (Tsubota et al. 2004).

Recent models have shifted to end-to-end (E2E) deep learning methods. One big advantage of E2E models is that it does not require forced alignment, pronunciation dictionary, or language models, which simplifies the complex process of traditional acoustic models. As the individual modules are integrated in a single network, the potential errors that were previously accumulated between the processes of building the ASR system are minimized as well. Most importantly, E2E models provide improved performances compared to the traditional acoustic models. In the APA domain, three different end-to-end architectures are experimented in Chen et al. (2018) to automatically learn acoustic and lexical cues. In Kyriakopoulos et al. (2018), an end-to-end Siamese network is used to compute phone distances. Leung et al. (2019) is the first study in the MDD domain to adopt E2E architecture, in particular CNN-RNN-CTC model. Lo et al. (2020) and Zhang et al. (2020) proposes a hybrid CTC/ATT, a hybrid usage of CTC and attention mechanism to detect and diagnose mispronunciations. These CNN-RNN-CTC and Hybrid CTC/ATT models are frequently compared in MDD studies as baselines (Algabri et al. 2022; Lin and Wang 2022; Peng et al. 2021; Wang et al. 2022).

With more large and high-end architectures proposed in various artificial

intelligence domains, the backbone acoustic models used in the CAPT systems are shifting rapidly to Transfomer and pre-trained self-supervised learning models as they show robustness in data scarce non-native assessments. Transformer (Vaswani et al. 2017) is a multi-head self-attention based encoder-decoder model that have shown the state-of-the-art performances in machine translation task and have since been the norm in natural language processing. In the APA task, Gong et al. (2022) have proposed Goodness-of-Pronunciation Transformer (GOPT) that takes the GOP measure as input to a Transformer-based architecture to assess multi-granular, multi-aspect scores. Developed versions of Transformer are also frequently introduced. Conformer (Gulati et al. 2020), a convolution augmented Transformer, was adopted in Fan et al. (2023) as the backbone architecture in individual APA and MDD models with separate task-specific layers and multi-width band. Squeezeformer (S. Kim et al. 2022), an optimized Conformer, is used as an encoder along with Transformer-based decoder in Guo et al. (2023) for MDD task.

2.1.1. Self-supervised learning model

Self-supervised learning (SSL) is a method in which a model learns general representations from unlabeled data. After the pre-training stage, the model is then fine-tuned with labeled data to meet the objective of each downstream task. Pre-trained with a vast amount of unlabeled audio data, SSL models have shown the state-of-the-art performances in various speech tasks including speech recognition and language identification with only a little amount of labeled data.

The pre-trained self-supervised learning models provide rich acoustic representation which can alleviate the data scarcity problem inherent in the CAPT

system and lead to better model performances. In the MDD task, Xu et al. (2021) and Peng et al. (2021) utilizes Wav2vec2.0 (Baevski et al. 2020) to fine-tune on CTC (Connectionist Temporal Classification) criteria. In the APA task, Chao et al. (2022) uses a mix of Wav2vec2.0, HuBERT, and WavLM to assess multi-granular, multi-aspect scores. Kim et al. (2022) compares different configurations of Wav2vec2.0 (Baevski et al. 2020; Hsu, Sriram, et al. 2021) and HuBERT (Hsu, Bolte, et al. 2021) to validate their usefulness.

2.1.1.1. Robust Wav2vec2.0

Wav2vec2.0 (Baevski et al. 2020) is the most representative model to utilize self-supervised learning scheme. As in Figure 1, the model consists of a multi-layer convolutional feature encoder, a quantization module, and a Transformer-based context network which is trained end-to-end. The output of the feature encoder z is separately fed into the quantization module and the context network, to respectively discretize the representation using Gumbel softmax objective, and learn the contextual information of the speech. A portion of latent speech representation is masked before being used as input for the context network. For each masked time step, the goal is to predict the correct quantized representation q for the context representation c among other quantized representation distractors using the contrastive loss L_m and the diversity loss L_d.

The pre-trained Wav2vec2.0 comes with two model sizes; BASE with 95m (approximated to 100m) parameters and LARGE with 317m (approximated to 300m) parameters. Both models are trained with recorded English audiobook dataset LibriSpeech (960 hrs.), and LARGE offers an additional version trained with

LibriVox (53.2k hrs.).

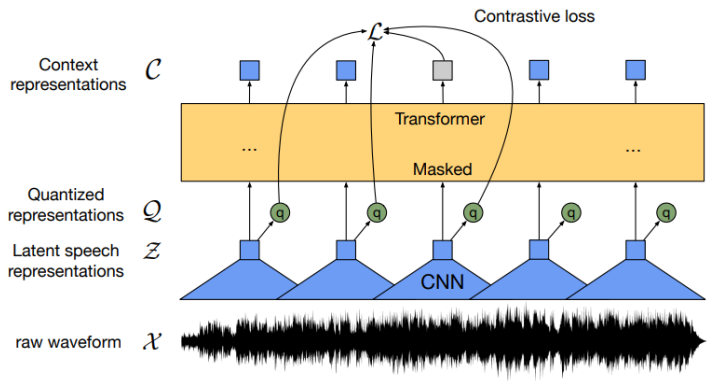

Figure 1 Illustration of the original Wav2vec2.0 from Baevski et al. (2020)

Robust Wav2vec2.0 (Hsu, Sriram, et al. 2021) BASE and LARGE share the same architecture from the original Wav2vec2.0, but are pre-trained with larger amount of unlabeled datasets from multiple domains to explore domain mismatch. To be specific, the LARGE model is pre-trained with 63k hours of English datasets from three domains, 60k hours of audiobook recordings (Libri-light), 2.3k hours of telephone conversation (Switchboard, Fisher), and 700 hours of Wikipedia recordings (Common Voice). Hsu et al. (2021) reveals that increasing the pre-trained data and domain variety helps models to achieve better speech recognition performances even for out-of-domain data, hence the name *robust*.

2.1.1.2. Wav2vec2.0 XLS-R

XLSR (Conneau et al. 2021) and the updated XLS-R (Babu et al. 2022) are multilingual versions of Wav2vec2.0 that were proposed to test cross-lingual performances for unsupervised pre-training. As illustrated in Figure 2, the models

are pre-trained on multiple languages at the same time, with the expectation that they will learn the discrete speech representations that are shared across languages. The training batches are sampled from multiple languages that form a multinominal distribution, with a parameter given to upsample low resource languages.

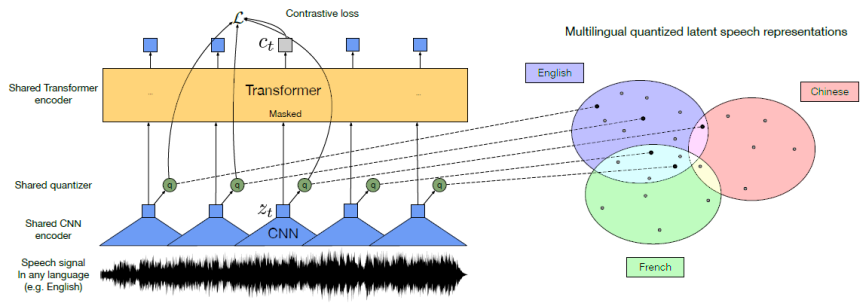

Figure 2 Illustration of XLSR from Conneau et al. (2021)

XLSR comes with BASE and LARGE models similar to Wav2vec2.0, whereas XLS-R comes with 300m, 1B, and 2B parameter sizes, making XLSR LARGE and XLS-R 300m to have the same architecture and size. The main difference between XLSR LARGE and XLS-R 300m lies in the amount of the data used for pre-training. The former is pre-trained with Common Voice, BABEL, and Multilingual Librispeech which consists a total 56k hours of 53 languages, whereas the latter is additionally pre-trained with VoxPopuli and VoxLingua107 datasets making up to 436k hours of 128 languages. Similar to the comparison for monolingual settings between Wav2vec2.0 and Robust Wav2vec2.0, Babu et al. (2022) reveal that the model with larger pre-training datasets show more robustness for out-of-domain datasets.

2.1.1.3. HuBERT

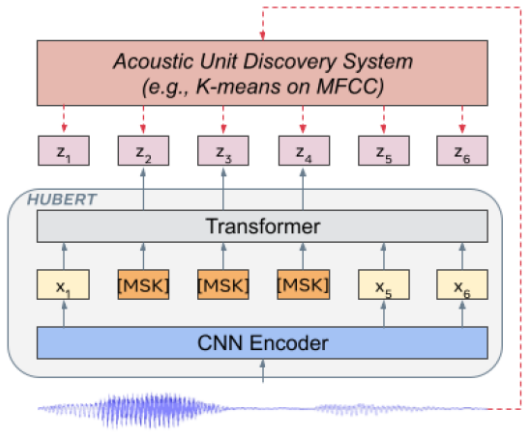

Figure 3 Illustration of HuBERT from Hsu, Bolte, et al. (2021)

HuBERT (Hsu, Bolte, et al. 2021) is another model to have shown state-of-the-art performances by utilizing self-supervised learning paradigm. As illustrated in Figure 3, the architecture extends Wav2vec2.0 with a convolution encoder and a Transformer network, and with inputs of the Transformer network being selectively masked. Unlike Wav2vec2.0 that adopts quantization module and a mix of contrastive loss and diversity loss, HuBERT (Hidden-Unit BERT) separates the quantization step and the masked prediction step and discovers *the hidden acoustic units* by iterative K-means clustering. Specifically, cross-entropy loss between the clustered units and outputs of the BERT network is computed to learn the discrete acoustic representation.

HuBERT is publicly available with three sizes, BASE, LARGE, and XLARGE, with BASE and LARGE having similar configurations to Wav2vec2.0 and XLARGE having 1B parameters. LARGE model is trained with 60k hours of

Libri-light dataset. BASE and LARGE/XLARGE models have different iteration methods. Whereas BASE starts from MFCC features clustering and then uses latent features of Transformer layer to subsequently cluster for a total of two clustering iterations, LARGE/XLARGE iterates only once, by leveraging the second iteration latent features of the BASE model.

2.1.1.4. WavLM

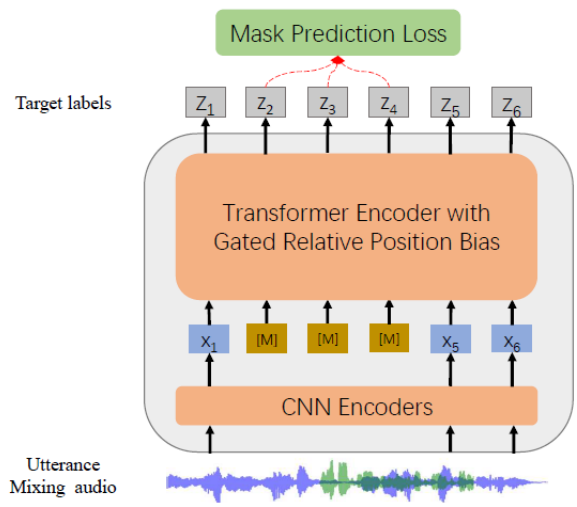

Figure 4 Illustration of WavLM from S. Chen et al. (2022)

WavLM (Chen et al. 2022) extends HuBERT and Wav2vec2.0. WavLM was proposed with the purpose of integrating various speech tasks other than ASR tasks to the usage of self-supervised learning speech models, as Wav2vec2.0 and HuBERT were tested with speech recognition and phone recognition tasks in their original work. To tackle this problem, WavLM implements masked speech denoising and prediction, and gated relative position bias as in Figure 4. WavLM comes with

three models, namely BASE and BASE + with 100m parameters and LARGE with 300m parameters similar to Wav2vec2.0 and HuBERT.

2.2. Acoustic features

Developing and selecting the appropriate acoustic features that embed non-native acoustic characteristics is another important research in APA and MDD literatures. If the acoustic models were to capture the generalized acoustic representation of speech, the acoustic features capture a specific quality of speech. The features can be extracted directly from speech using respective formula, such as fundamental frequency and energy (van Dalen, Knill, and Gales 2015). Other features depend on speech recognized from acoustic models, as in the case of silence duration, number of words per second, and articulation rate (Ryu et al. 2016; Zechner et al. 2009).

2.2.1. Goodness-of-Pronunciation measure

The Goodness-of-Pronunciation (GOP) measure also belongs to the traditional feature-based methods, and is one of the most representative features used for pronunciation assessments and mispronunciation detection and diagnosis for its high correlation with human experts. GOP measure is a duration normalized log posterior probability of phone p, as defined below from the original paper of Witt and Young (2000):

$$GOP(\mathrm{p}) = \frac{\left|\log\left(P(p|O^p)\right)\right|}{NF(p)}$$

$$= \frac{\left|\log\left(\frac{P(O^p|p)P(p)}{\sum_{q\in Q} p(O^p|q)P(q)}\right)\right|}{NF(p)}$$

$$= \frac{\left|\log\left(\frac{P(O^p|p)}{\max_{q\in Q} p(O^p|q)}\right)\right|}{NF(p)}$$

where $P(p|O^p)$ is the posterior probability of a speaker to have uttered phone p given the corresponding audio segment O^p, and $NF(p)$ is the number of the frames in the segment O^p. The posterior probability $P(p|O^p)$ can be expressed using (1) the likelihood $p(O^p|q)$ for each phone q to match the audio segment O^p, (2) and the prior probabilities $P(p)$ and $P(q)$, where Q is the set of all possible phones corresponding to pronunciation O^p. This is under the assumption that all phones have equal probability, $P(p) = P(q)$ and that the sum of all possible phones can be approximated to maximum.

Similar to other acoustic model-based features, the GOP formula needs acoustic model to compute probabilities. Thus on one side, studies worked on adopting better acoustic models that conform well to GOP. In APA domain, the original paper Witt and Young (2000) extract GOP from GMM-HMM for phone-level scoring. Tong et al. (2014) use Subspace GMM (SGMM) for fluency scoring. In the MDD domain, Hu et al. (2015) also extend GMM-HMM to DNN-HMM for better acoustic model on GOP measure. On the other side, studies have attempted to develop the GOP measure itself to achieve higher performances for the target task. In the APA domain, Sudhakara et al. (2019) and Lin and Wang (2021a) each present noise-robust GOP to cope with assessments in practical education settings.

Sudhakara, Ramanathi, Yarra, and Ghosh (2019) derive a GOP formula that considers senone posterior probability and HMM state transition probability. In the MDD domain, Ryu and Chung (2017) applies articulatory features to GOP (aGOP) to provide more sophisticated corrective feedback. Shi et al. (2020) similarly propose a context-aware GOP (CaGOP) that uses transition factor and duration factor to account for frame-level phone transition in mispronunciation detection.

Although recent MDD works have seen transitions to end-to-end speech recognition-based models, GOP still remains as baseline (E. Kim et al. 2022; Lin and Wang 2021b, 2022) and as strong input in state-of-the-art models (Chao et al. 2022; Gong et al. 2022) in APA task.

2.3. The limitation of APA and MDD works

However, few studies have attempted to integrate the two tasks regardless of their similarities in methodology. One reason lies in the limited annotations provided in the datasets frequently used for CAPT studies. L2-ARCTIC (Zhao et al. 2018), a dataset frequently used to measure MDD performance, only provides canonical phone sequences and the realized phone sequences with no pronunciation score labels. Speechocean762 (Zhang et al. 2021) is a dataset that covers both APA and MDD tasks, with detailed scores of multi-aspect, multi-granularity and mispronunciation annotations. However, as the dataset was made with the aim of assessing pronunciations, most literatures focus on APA task for Speechocean762 (Gong et al. 2022; E. Kim et al. 2022). Few studies use Speechocean762 as an MDD benchmark. D. Zhang et al. (2022) show their data augmentation method improves MDD performance on out-of-domain Speechocean762 test set, and Wadud et al.

(2023) adopt a non-autoregressive framework for MDD. Nonetheless, they also focus only on the MDD task and do not give clear details on how the data was preprocessed for the task.

A few studies that mention both tasks still regard one as auxiliary or separate. Lin et al. (2020) cover mispronunciation detection as an auxiliary, binary phone-level scoring in multi-granular APA. Lin & Wang (2023) utilize native(-like) data and the matching canonical phones for auxiliary CTC training to assist holistic/accuracy APA. Zheng et al. (2022) perform phone-level pronunciation assessment along with MDD, but as two separate tasks that can be achieved with respective APA and MDD datasets for fine-tuning. In a similar sense, Fan et al. (2023) adopt Conformer to evaluate on APA and MDD datasets, but with separate models of different task-specific layers.

Given the methodological similarity and the high linguistic correlation between automatic pronunciation assessment and mispronunciation detection and diagnosis task, this study proposes to integrate the two tasks in a joint model.

Chapter 3. Methodology

3.1. Proposed method

The training process of the proposed joint model for pronunciation assessment and mispronunciation detection and diagnosis is illustrated in Figure 5. First, **pre-trained self-supervised learning model** is employed as the backbone model for the joint model. Second, the SSL model is additionally fine-tuned to make **phone recognition model**. This is to test which of the two backbone models transfer well to bring better assessment performances for the joint model. Lastly, **joint model** of APA and MDD is trained using multi-task learning.

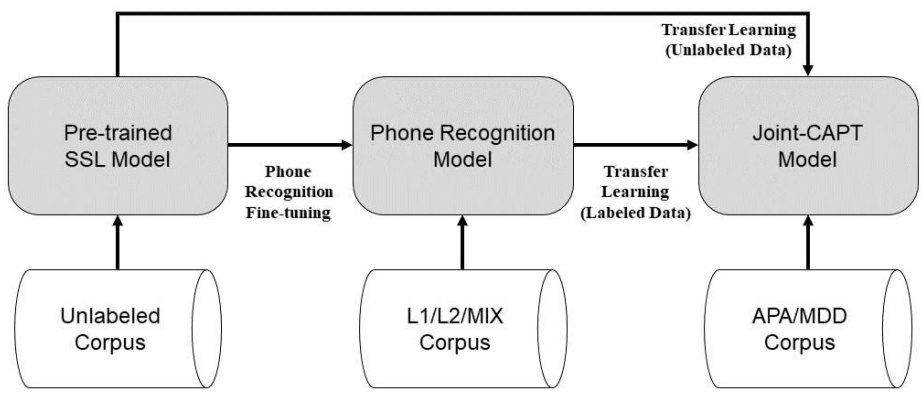

Figure 5 The training process of the proposed method

Thus the three steps can be categorized into **(1)** transfer learning with an auxiliary fine-tuning on phone recognition, and the main **(2)** multi-task learning of APA and MDD. Chapter 3.1.1, Chapter 3.1.2, and Chapter 3.1.3 explains in detail the architecture of each steps.

3.1.1. Pre-trained self-supervised learning model

Transfer learning (TL) takes a resource-rich, huge model from another domain to adapt to the target domain. As non-native speech suffers from the inherent problem of data scarcity, the usage of SSL as backbone architecture has been a frequent practice in the CAPT domain as mentioned in Chapter 2.1.1. Following previous studies, this work uses SSL model as backbone acoustic model to transfer its robust speech representation. In this paper, four different SSL models are explored, Robust Wav2vec2.0 (Hsu, Sriram, et al. 2021), Wav2vec2.0 XLS-R (Babu et al. 2021), HuBERT (Hsu, Bolte, et al. 2021), and WavLM (Chen et al. 2022).

Wav2vec2.0 (Baevski et al. 2020) and its multi-lingual version XLSR (Conneau et al. 2021) have shown competitive performances in CAPT literatures of APA and MDD tasks. This work uses Robust Wav2vec2.0 and XLS-R with bigger amount of pre-training dataset as the original works showed the models to have more robustness. HuBERT and WavLM are also experimented as they have been relatively underexplored in the CAPT domain. To see the effect of different datasets and learning schemes, the size of all the models was controlled to 300 million parameters, which is the biggest size provided for Wav2vec2.0. Accordingly, LARGE robust model, XLS-R 300m model, LARGE HuBERT model, and LARGE WavLM model were used.

3.1.2. Phone recognition model

This work additionally fine-tunes the SSL model on phone recognition with different speech characteristics before multi-task learning: a native dataset, a non-native

dataset, and a sum of the two datasets. This is to see if the extra fine-tuning can enhance the model with better acoustic representation compared to raw self-supervised learning model, and if so, to see what types of speech data help when transferred on the joint model. For fine-tuning, a fully-connected layer (language model head) is added on top of the Transformer network of the SSL model to train on Connectionist Temporal Classification (CTC) loss.

3.1.3. Joint model

Multi-task learning (MTL) simultaneously trains tasks with different objective functions using a shared model. With the increased information, downstream speech tasks including emotion recognition (Cai et al. 2021) or dysarthria assessment (Yeo et al. 2023) have leveraged the framework to gain more generalized models. Motivated by its effectiveness in various speech domains, this work utilizes multi-task learning to jointly train APA and MDD. With the joint optimization, the model is expected to learn the correlation between the output pronunciation scores and phone sequences to gain performance increases than respective single-task learning (STL).

Three architectures are suggested for the joint model, shown in Figure 6 and 7. The difference of each architecture lies in the APA training. All architectures utilize the SSL encoder and its weights. For the raw audio input $x \in \mathbb{R}^L$ with length L, the SSL encoder outputs T sequences of D dimensional latent speech representation $h \in \mathbb{R}^{T \times D}$. For the MDD task, the latent speech representation h passes the same fully connected layer used in phone recognition fine-tuning to leverage the fine-tuned weights. The output logit $\hat{z} \in \mathbb{R}^{T \times V}$ is optimized using CTC

loss (L_{MDD}) with the ground-truth realized phone annotations after a softmax operation, where V is the size of the vocabulary.

SSL-based joint model

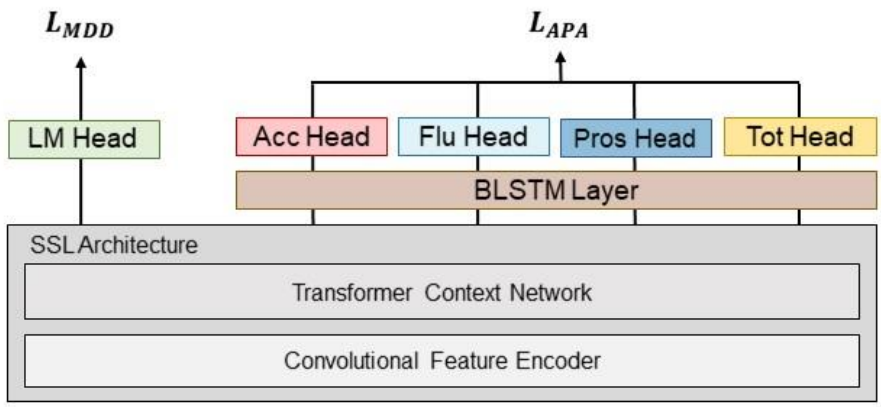

Figure 6 The architecture of the joint model (CE/RMSE)

The first and second architecture are entirely based on SSL acoustic model as shown in Figure 6. For the APA task, the latent speech representation goes through an additional bidirectional long short-term memory (BLSTM) layer shared among four pronunciation assessment tasks to capture the information shared between assessments. The model yields $\bar{h} \in \mathbb{R}^{T \times H}$ where H is the size of the output hidden dimension. The output representation is then passed to each assessment head which consists of a fully connected layer and an average pooling over time dimension.

The difference between the first and the second architecture lies in the APA loss. The first architecture uses cross-entropy loss (**CE**), thus the output makes $\hat{y}\{\text{acc}, \text{flu}, \text{pros}, \text{tot}\} \in \mathbb{R}^{C}$, which are logits of accuracy, fluency, prosodic, and total

score, respectively where C is the number of labels. Logits of each aspect are stacked to make final APA logits $\hat{y} \in \mathbb{R}^{C\times 4}$ to be optimized using cross-entropy criteria (L_{APA}) with the ground-truth score labels after a softmax operation. The second architecture uses root mean square error loss (**RMSE**), thus the output makes $\hat{y}\{\text{acc}, \text{flu}, \text{pros}, \text{tot}\} \in \mathbb{R}^1$, which are logits of accuracy, fluency, prosodic, and total score. Logits of each aspect are concatenated to make final APA logits $\hat{y} \in \mathbb{R}^4$ to be optimized using RMSE criteria (L_{APA}) with the ground-truth score labels.

SSL-based joint model with additional GOP input

As demonstrated in Chapter 2.2.1, Goodness-of-Pronunciation measure remains competitive in APA task. Thus the usefulness of GOP on the joint model is tested by giving GOP measure as additional input for the third architecture. The APA loss is fixed to RMSE (**RMSE+GOP**). GOP measure is extracted following the formulae from Zhang et al. (2021) and Gong et al. (2022). To extract the feature, the audio and the matching canonical phone transcription of the target dataset with a set of 42 phones is given as input to a trained acoustic model. The acoustic model is a publicly available model that is based on the factorized time-delay neural network (TDNN-F) and trained using the Kaldi Librispeech S5 recipe (Povey et al., 2011). Force-aligned at the phone-level, the output consists of 84-dimensional GOP measure.

Figure 7 shows its architectural difference from the first two architectures. For the APA task, the latent speech representation goes through an additional bidirectional long short-term memory (BLSTM) layer shared among four pronunciation assessment tasks to capture the information shared between

assessments, followed by a pooling over time dimension. On the other side, the extracted GOP measure also goes through a BLSTM layer to be pooled over time dimension. The two pooled features are then added, to yield $\bar{h} \in \mathbb{R}^H$ where H is the size of the output hidden dimension. The output representation is then passed to each assessment head, a fully connected layer, to make logits of accuracy, fluency, prosodic, and total score. The logits are concatenated to make final APA logits $\hat{y} \in \mathbb{R}^4$ to be optimized using RMSE loss (L_{APA}) with the ground-truth score labels.

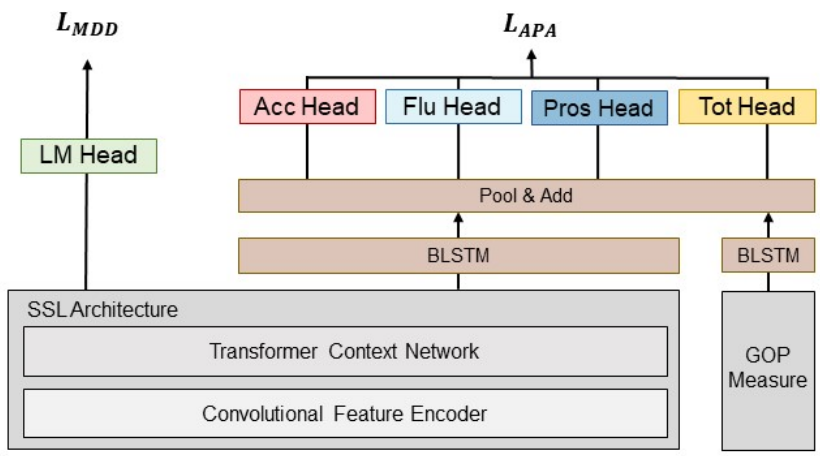

Figure 7 The architecture of the joint model (RMSE+GOP)

The classification heads and the language model head are then optimized using the joint loss L_{CAPT} for all three architectures, which is a combination of L_{APA} and L_{MDD}:

$$L_{CAPT} = \alpha L_{APA} + \beta L_{MDD}$$

where α and β are used to balance the two losses. α is chosen from the set of $\alpha \in$

$\{0.05, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$ and β is fixed to 1.0. This is to adjust the weights on L_{APA} as L_{MDD} optimizes faster when given auxiliary fine-tuning. This enables fair comparison between models using raw SSL acoustic models and fine-tuned models.

3.2. Experiment settings

3.2.1. Datasets

Experiments are conducted using three public datasets. For the auxiliary fine-tuning on phone recognition, TIMIT (Garofolo et al. 1993) and L2-ARCTIC (Zhao et al. 2018) are used, with each dataset representing phonetically labeled data of native (L1) and non-native (L2) speech. For the main task of joint APA and MDD, Speechocean762 (Zhang et al. 2021) is used. The summary of each dataset is represented in Table 1.

Table 1 Summary of the datasets used for experiments

	TIMIT	**L2-ARCTIC**	**Speechocean762**	
Split	Train	Train	Train	Test
Hours	3.94	2.79	2.88	2.69
Utterances	4620	2699	2500	2500
Data Type (L1)	Native (English)	Non-native (Hindi, Korean, Mandarin, Spanish, Arabic, Vietnamese)	Non-native (Mandarin)	

TIMIT is a native speech dataset that contains recordings of 8 US English dialects and is phonetically transcribed with 61 phone set. The original TIMIT train

split is used for the fine-tuning. L2-ARCTIC version 5.0 is a non-native speech dataset that contains English of six L1 backgrounds, including Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese, and is transcribed with 40 phone set. For fine-tuning, the suggested L2-ARCTIC train split from Feng et al. (2020) and Zheng et al. (2022).

Speechocean762 is another non-native speech dataset that contains English read speech of Mandarin speakers. Designed to support pronunciation assessment studies, the dataset provides rich labels of multi-granular, multi-aspect scores for each utterance. Five experts independently assessed the scores independently. For phone-level score, accuracy score is provided with a range of 0-2. For word-level scores, accuracy, stress, and total scores are provided and for sentence-level scores, accuracy, completeness, fluency, prosodic, and total scores are provided with a range of 0-10 for both granularities. Accuracy scores are used to detect mispronunciations or heavy-accents in each granularity. Completeness score reflects the amount of words that are realized. Fluency score is used to assess the amount of pauses, repetition, or stammering in the utterance. Prosodic score measures intonation, speed, and rhythm. Stress score is used to indicate correct stress position or mono-syllabic word (and is measured into binary scores of 5 or 10). The scores of higher granularity does not mean a simple average of lower granularity and are individually measured, such as in accuracy and total scores in word-level and sentence-level. However, if the gap between the scores of different granularities seemed unreasonable, the labelling platform gave a warning to the experts.

Speechocean762 provides a canonical phone transcription for the text script, along with an extra mispronunciation transcription for inaccurate phones. The

canonical phones are transcribed using 39 phones following CMUDict (Carnegie Mellon University, 2000). The mispronounced phones are transcribed using 46 phone set, the same 39 phones from CMUDict, <unk> for unknown phones, and six L2 phones ‘AR’, ‘DR’, ‘DZ’, ‘IR’, ‘TR’, ‘TS’. Mispronunciations take up to 4% of the train set and 3% of the test set in the phone transcriptions. Out of the mispronunciations, <unk> takes up to 26% and 25%, respectively.

In this work, the original train and test split with the unified labels between the experts was used. For APA task, four aspects of sentence scores were used, namely accuracy, fluency, prosodic, and total. For MDD task, realized phone transcription is used which was created by replacing the canonical phone transcription with mispronunciation transcription for inaccurate phones. Thus the final phone set equals to a total of 46 phones. The same phone set was applied to the auxiliary phone recognition fine-tuning. The phone sets of TIMIT and L2-ARCTIC were mapped into CMUDict to be combined with Speechocean762 phone set and this 46 phone set was used.

3.2.2. Evaluation metrics

The APA performance is measured using Pearson Correlation Coefficient (PCC) between model prediction scores and human annotated scores. For MDD performance, Precision, Recall, and F1 scores are computed according to the metrics used in Li et al. (2017) and Leung et al. (2019) for both correct pronunciations and mispronunciations following Wadud et al. (2023). True Acceptance (TA) refers to predicting correct pronunciation as correct pronunciations, False Acceptance refers to predicting mispronunciations as correct pronunciations, True Rejection (TR)

refers to predicting mispronunciations as mispronunciations, and False Rejection (FR) refers to predicting correct pronunciations as mispronunciations. As metrics are reported for both correct pronunciations and mispronunciations, classes used for metrics contradict each other. True Positive (TP) corresponds to True Acceptance for correct pronunciations and True Rejection for mispronunciations. False Positive (FP) corresponds to False Acceptance for correct pronunciations and False Rejection for mispronunciations. True Negative (TN) corresponds to True Rejection for correct pronunciations and True Acceptance for mispronunciations. False Negative (FN) corresponds to False Rejection for correct pronunciations and False Acceptance for mispronunciations. Using the definitions, Precision, Recall and F1 metric each follows the following formula:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.2.3. Implementation and experiment details

For all architectures, models had the feature encoder frozen and were trained with 8 batch sizes, an AdamW optimizer, a training epoch of 100, and a linear scheduler with a learning rate of 1e-4 and a warm-up ratio of 0.1. Batches for CE models were sampled by audio length. Batches for RMSE and RMSE+GOP models were sampled by giving more weights to imbalanced labels of accuracy score. Pre-trained SSL models were implemented using HuggingFace (Wolf et al. 2020). All the

experiments were repeated for 3 trials with different random seeds and are reported with the mean and standard deviation value. BLSTM layers were fixed to 128 hidden dimensions.

For the experiment results of shown in Chapter 4, Robust Wav2vec2.0 is used as the base backbone model, α is set to 0.25 for the base multi-task learning loss weight. The performance of the joint model will be first compared to respective APA and MDD models, for the three suggested architectures (**CE, RMSE, RMSE+GOP**). Results will be presented for both raw self-supervised learning model **(SSL)** and models fine-tuned with native dataset TIMIT (**L1**), for leveraging TIMIT as additional train dataset or transfer learning is a common practice in CAPT studies. Then for the best architecture, the effects of using fine-tuned model on phone recognition for the backbone model are more deeply explored (L2-ARCTIC (**L2**), a sum of TIMIT/L2-ARCTIC (**MIX**)). In a similar sense, the effects of using different backbone SSL models (**XLS-R**, **HuBERT**, **WavLM**) are also compared.

Chapter 4. Results

Chapter 4 demonstrates the experiment results with regard to transfer learning, and multi-task learning of automatic pronunciation assessment and mispronunciation detection and diagnosis tasks. In Chapter 4.1, the results of the three architectures (**CE**, **RMSE**, **RMSE+GOP**) for the proposed joint model (**Joint-CAPT** [$\alpha = 0.25$, $\beta = 1.0$]) are presented. They are each compared to respective single-task learning (**APA** [$\alpha = 1.0$, $\beta = 0.0$], **MDD** [$\alpha = 0.0$, $\beta = 1.0$]), for the baseline raw Robust Wav2vec2.0 (**SSL**) and its fine-tuned model with native dataset (**L1**). Chapter 4.2 and Chapter 4.3 compare the effects of auxiliary phone recognition fine-tuning (**SSL, L1, L2, MIX**) and the backbone self-supervised learning model (**Robust, XLS-R, HuBERT, WavLM**) on the joint model for the architecture that showed the best performance. The results are presented in separate tables for APA and MDD, although the joint models simultaneously learn and infer the results.

4.1. Comparison of results on APA, MDD, and joint models of different architectures

Table 2, Table 3, and Table 4 present the APA results for the proposed Joint-CAPT with cross-entropy APA loss, root mean square error APA loss, and additional Goodness-of-Pronunciation measure input. The results are compared to those of respective single-task APA model.

Table 2 Experiment results for APA task with regard to multi-task learning for the

architecture CE

	Model	Pronunciation Scores (PCC)			
		Accuracy	Fluency	Prosodic	Total
CE	APA-SSL	0.609 ±0.036	0.652 ±0.026	0.650 ±0.034	0.633 ±0.024
	Joint-CAPT-SSL	0.714 ±0.003	0.763 ±0.006	0.767 ±0.004	0.732 ±0.005
	APA-L1	0.629 ±0.065	0.738 ±0.029	0.733 ±0.036	0.680 ±0.041
	Joint-CAPT-L1	**0.719 ±0.005**	**0.775 ±0.000**	**0.773 ±0.006**	**0.743 ±0.010**

Table 2 shows that multi-task learning greatly improves pronunciation assessment performance for the architecture with cross-entropy criteria chosen as APA loss. Joint-CAPT-SSL and Joint-CAPT-L1 both have higher PCC for all scores than the respective APA-SSL and APA-L1, with an average of 0.108 and 0.057 increase. The average PCC of Joint-CAPT-SSL is even higher than APA-L1 which leverages the extra knowledge by a mean of 0.049. Auxiliary phone recognition fine-tuning also improves APA performance, with an average increase of 0.059 between APA-SSL and APA-L1, and an average increase of 0.008 between Joint-CAPT-SSL and Joint-CAPT-L1.

Table 3 Experiment results for APA task with regard to multi-task learning for the architecture RMSE

	Model	Pronunciation Scores (PCC)			
		Accuracy	Fluency	Prosodic	Total
RMSE	APA-SSL	0.687 ±0.003	0.763 ±0.006	0.770 ±0.003	0.713 ±0.003
	Joint-CAPT-SSL	0.722	**0.797**	**0.795**	0.740

		±0.008	**±0.002**	**±0.003**	±0.006
	APA-L1	0.680 ±0.007	0.769 ±0.013	0.771 ±0.003	0.714 ±0.005
	Joint-CAPT-L1	**0.730 ±0.001**	**0.797 ±0.003**	0.792 ±0.003	**0.746 ±0.001**

Table 3 shows that multi-task learning also greatly improves pronunciation assessment performance for the architecture that uses RMSE as the APA loss function, with Joint-CAPT-SSL and Joint-CAPT-L1 having higher PCC for all scores than the counterpart APA-SSL and APA-L1. The respective pairs showed an average gap of 0.031 and 0.032. However unlike with cross-entropy loss, auxiliary phone recognition fine-tuning showed conflicting results for APA performance. Although Joint-CAPT-L1 showed a slight performance improvement of 0.002 from Joint-CAPT-SSL on average, the joint model which used raw Robust Wav2vec2.0 as backbone model showed better correlation with human evaluators for prosodic evaluation. This was the same for APA-L1 that had lower correlation with human evaluations for accuracy score than APA-SSL.

Table 4 Experiment results for APA task with regard to multi-task learning for the architecture RMSE+GOP

Model		Pronunciation Scores (PCC)			
		Accuracy	Fluency	Prosodic	Total
RMSE+GOP	APA-SSL	0.698 ±0.002	0.778 ±0.005	0.778 ±0.005	0.724 ±0.003
	Joint-CAPT-SSL	**0.751 ±0.004**	**0.815 ±0.002**	**0.810 ±0.002**	**0.768 ±0.002**
	APA-L1	0.705 ±0.001	0.784 ±0.001	0.782 ±0.002	0.729 ±0.009
	Joint-CAPT-L1	0.741	0.805	0.803	0.756

	±0.004	±0.004	±0.001	±0.004

Table 4 shows the APA results for the architecture with RMSE loss and additional GOP input. Similar to the previous two architectures, multi-task learning greatly improves pronunciation assessment performance. Joint-CAPT-SSL and Joint-CAPT-L1 respectively showed an average correlation of 0.786 and 0.776, which was higher than the average correlation of the respective models APA-SSL and APA-L1 that each showed 0.745 and 0.750. This not only applied to the average but to all four scores as well. However, auxiliary phone recognition fine-tuning worsened the APA performance for the joint model with Joint-CAPT-SSL achieving higher correlations than Joint-CAPT-L1 on all aspects. However, for single-task learning APA model, leveraging the extra speech representation of native speakers was useful as shown in the comparison between APA-SSL and APA-L1.

To sum up, multi-task learning improved the APA performance on four aspects for all three architectures. This also applied to both cases where the backbone model was either raw Robust Wav2vec2.0 or its fine-tuned model with labeled native dataset. However unlike multi-task learning, the auxiliary phone recognition fine-tuning showed conflicting results, with joint model of RMSE+GOP showing worse performance with fine-tuned backbone model. Between the different joint model architectures presented in Table 2, 3, and 4, the APA performance was the highest for the architecture with RMSE loss and additional GOP measure, followed by the architecture with RMSE loss, and the architecture with cross-entropy loss, for all four aspects. Again, this tendency was the same for both cases of SSL and L1 backbone model. In total, the correlation of the Joint-CAPT-SSL with RMSE loss

and GOP measure showed the best performance for all four scores, as marked bold in Table 4.

Table 5, Table 6, and Table 7 each presents the MDD results for the proposed Joint-CAPT with cross-entropy APA loss, root mean square error APA loss, and additional Goodness-of-Pronunciation measure input. The results are compared to respective single-task MDD models. As the architecture difference lies in APA task, single-task MDD models (MDD-SSL, MDD-L1) have the same architecture for all CE, RMSE, and RMSE+GOP and thus should show similar results. However, due to the different sampling methods between CE and RMSE/RMSE+GOP, there are result differences between the two groups.

Table 5 Experiment results for MDD task with regard to multi-task learning for the architecture CE

Model		PER (%)	Correct Pronunciations			Mispronunciations		
			Prec.	Recall	F1	Prec.	Recall	F1
CE	MDD-SSL	**9.89** ±0.024	0.997 ±0.000	0.928 ±0.000	0.961 ±0.000	0.267 ±0.001	0.914 ±0.005	0.413 ±0.002
	Joint-CAPT-SSL	9.91 ±0.030	0.997 ±0.000	**0.929** ±0.001	**0.962** ±0.000	**0.268** ±0.003	0.914 ±0.002	**0.415** ±0.003
	MDD-L1	9.90 ±0.041	0.997 ±0.000	0.927 ±0.001	0.961 ±0.000	0.265 ±0.002	**0.916** **±0.003**	0.410 ±0.003
	Joint-CAPT-L1	9.93 ±0.068	0.997 ±0.000	0.928 ±0.001	**0.962** ±0.000	0.267 ±0.002	0.914 ±0.004	0.414 ±0.003

Table 5 shows that although more subtle than APA performance increase, multi-task learning also improves MDD performance. For both correct pronunciations and mispronunciations, Joint-CAPT-SSL (0.962, 0.415) and Joint-CAPT-L1 (0.962, 0.414) had higher F1 scores of than the respective MDD-SSL and MDD-L1. The performance gain was achieved from higher recall for correct

pronunciations, and higher precision for mispronunciations. However, auxiliary phone recognition slightly reduces MDD performance as F1 scores were reduced for both MDD-L1 and Joint-CAPT-L1, caused by lower precision of mispronunciations.

Table 6 Experiment results for MDD task with regard to multi-task learning for the architecture RMSE

Model		PER (%)	Correct Pronunciations			Mispronunciations		
			Prec.	Recall	F1	Prec.	Recall	F1
RMSE	MDD-SSL	**10.16 ±0.001**	0.997 ±0.000	**0.927 ±0.000**	**0.961 ±0.000**	0.264 ±0.001	0.915 ±0.003	0.410 ±0.002
	Joint-CAPT-SSL	10.24 ±0.001	0.997 ±0.000	**0.927 ±0.000**	**0.961 ±0.000**	**0.265 ±0.000**	0.917 ±0.006	**0.411 ±0.001**
	MDD-L1	10.36 ±0.001	0.997 ±0.000	0.924 ±0.001	0.959 ±0.000	0.256 ±0.003	0.915 ±0.006	0.400 ±0.004
	Joint-CAPT-L1	10.34 ±0.001	**0.998 ±0.000**	0.925 ±0.000	0.960 ±0.000	0.261 ±0.001	**0.923 ±0.004**	0.406 ±0.002

Table 6 shows that multi-task learning is also valid when used with RMSE loss. Joint-CAPT-SSL and Joint-CAPT-L1 had higher F1 scores than the respective MDD-SSL and MDD-L1. In the same manner, auxiliary phone recognition reduces MDD performance for mispronunciations, with F1 scores reduced for both MDD-L1 and Joint-CAPT-L1.

Table 7 Experiment results for MDD task with regard to multi-task learning for the architecture RMSE+GOP

Model		PER (%)	Correct Pronunciations			Mispronunciations		
			Prec.	Recall	F1	Prec.	Recall	F1
RMSE +	MDD-SSL	**10.16 ±0.001**	0.997 ±0.000	0.927 ±0.000	**0.961 ±0.000**	0.264 ±0.001	0.915 ±0.003	0.410 ±0.002
	Joint-	10.19	0.997	**0.928**	**0.961**	**0.267**	0.918	**0.414**

GOP[1]	CAPT-SSL	±0.001	±0.000	**±0.000**	**±0.000**	**±0.001**	±0.002	**±0.002**
	MDD-L1	10.36 ±0.001	0.997 ±0.000	0.924 ±0.001	0.959 ±0.000	0.256 ±0.003	0.915 ±0.006	0.400 ±0.004
	Joint-CAPT-L1	10.38 ±0.000	**0.998** ±0.000	0.925 ±0.000	0.960 ±0.000	0.259 ±0.000	**0.922** **±0.004**	0.405 ±0.000

In line with the results from Table 5 and 6, Table 7 shows that multi-task learning also improves mispronunciation detection and diagnosis performance when used with additional GOP measure. Again, this applied to both SSL and L1 models. Joint-CAPT-SSL showed F1 scores of 0.414 for mispronunciations which was higher than 0.411 of MDD-SSL, and Joint-CAPT-L1 showed F1 scores of 0.405 which was higher than the scores of MDD-L1 of 0.400. The precision and recall was improved in general for both correct pronunciations and mispronunciations. However, MDD-SSL and Joint-CAPT-SSL showed decreased F1 scores on mispronunciations when transferred from models with additional native knowledge. This is in accordance with the results for APA task which showed that additional phone recognition fine-tuning does not help when using additional GOP features.

To sum up, multi-task learning improved the MDD performance for either raw Robust Wav2vec2.0 backbone model or its fine-tuned model on labeled native dataset. Unlike multi-task learning, the auxiliary phone recognition deteriorated the MDD performance on both single-task and joint models for all CE, RMSE, RMSE+GOP architectures. Between the different architectures presented in Table 5, 6, and 7, models with RMSE loss and weighted sampling showed an increase of PER and also a slight decrease in MDD performances compared to the models with cross-

[1] The results from Table 6 are shown for MDD-SSL and MDD-L1 as they have the same architecture and configuration.

entropy loss.

Altogether, the joint model showed the highest performance in both pronunciation assessment and mispronunciation detection and diagnosis tasks for all experimented architectures. This proves the effectiveness of jointly training APA and MDD tasks. In particular, Joint-CAPT-SSL model with RMSE loss and additional GOP features showed the best results for APA task and competitive results for MDD task. Thus for the following chapters, Joint-CAPT-SSL of RMSE+GOP is used as the baseline for backbone model comparison and results analysis.

4.2. Comparison of results on auxiliary phone recognition

Table 8 Experiment results APA task with regard to using a fine-tuned model as backbone architecture for the joint model (RMSE+GOP)

Auxiliary fine-tuning	Pronunciation Scores (PCC)			
	Accuracy	Fluency	Prosodic	Total
SSL (baseline)	**0.751 ±0.004**	**0.815 ±0.002**	**0.810 ±0.002**	**0.768 ±0.002**
L1	0.741 ±0.004	0.805 ±0.004	0.803 ±0.001	0.756 ±0.004
L2	0.735 ±0.002	0.808 ±0.004	0.801 ±0.002	0.755 ±0.005
MIX	0.732 ±0.002	0.809 ±0.006	0.803 ±0.005	0.752 ±0.003

The effects of using a fine-tuned model as backbone architecture is presented in Table 8 and 9. The raw Robust Wav2vec2.0 is the baseline. For APA task, transferring knowledge from models with auxiliary phone recognition fine-tuning did not help the joint model to achieve better correlations. Joint-CAPT-SSL achieved higher

correlations than Joint-CAPT-L1, Joint-CAPT-L2 and Joint-CAPT-MIX. Between the models that leverage auxiliary fine-tuning, although the average correlation of Joint-CAPT-L1 is slightly higher than Joint-CAPT-L2 and Joint-CAPT-MIX, each individual score shows conflicting results. This is an interesting finding given that L2 has more similar acoustic characteristics to the target Speechocean762 data, and the mix of the two has a larger amount of data.

Table 9 Experiment results for MDD task with regard to using a fine-tuned model as backbone architecture for the joint model (RMSE+GOP)

Auxiliary fine-tuning	PER (%)	Correct Pronunciations			Mispronunciations		
		Prec.	Recall	F1	Prec.	Recall	F1
SSL (baseline)	**10.19 ±0.001**	0.997 ±0.000	**0.928 ±0.000**	**0.961 ±0.000**	**0.267 ±0.001**	0.918 ±0.002	**0.414 ±0.002**
L1	10.38 ±0.000	**0.998** ±0.000	0.925 ±0.000	0.960 ±0.000	0.259 ±0.000	**0.922 ±0.004**	0.405 ±0.000
L2	10.46 ±0.000	0.997 ±0.000	0.924 ±0.001	0.959 ±0.000	0.257 ±0.002	0.918 ±0.003	0.401 ±0.003
MIX	10.37 ±0.000	**0.998** ±0.000	0.924 ±0.000	0.960 ±0.000	0.259 ±0.000	0.921 ±0.003	0.404 ±0.000

Auxiliary phone recognition fine-tuning did not aid MDD task as well. Joint-CAPT-SSL showed the highest F1 scores on correct pronunciations and mispronunciations when compared to Joint-CAPT-L1, Joint-CAPT-L2, and Joint-CAPT-MIX that transferred from models with additional native and non-native knowledge. The latter three models had lower recall for correct pronunciations, and lower precision for mispronunciations.

In all, transfer learning from auxiliary phone recognition fine-tuning did not help, with Joint-CAPT-SSL showing the highest performance in both pronunciation

assessment and mispronunciation detection and diagnosis tasks.

4.3. Comparison of results on self-supervised learning model

The effect of using different SSL models as backbone architecture is presented in Table 10 and 11. Robust Wav2vec2.0 is the baseline, and all the experiments use raw SSL models. XLS-R consistently showed the weakest performance in both APA and MDD tasks. The correlation with human experts was the lowest for all four scores, with an average of 0.761. The F1 scores of correct pronunciations and mispronunciations were also visibly lower than other three models. Although XLS-R has been pre-trained with the biggest amount of audio, the multi-lingual characteristics of the acoustic embeddings may not have been fit for the scope as the target dataset is fixed to English. This is similar to the results of models that were transferred from phone recognition models trained with L2 and MIX datasets. As to how the transferring L2 dataset did not help non-native speech assessment, multi-lingual pre-training did not aid model performances although they contain diverse phonetic patterns including the speech of native Mandarin and thus may share more similar acoustic characteristics to the target Speechocean762 data.

HuBERT showed similar performances to the baseline Robust Wav2vec2.0, with higher correlation for total score in APA task and higher F1 score for correct pronunciations in MDD task than Robust Wav2vec2.0. However on average, HuBERT showed an average correlation of 0.783 which was slightly lower than the average correlation of 0.786 of Robust Wav2vec2.0. The F1 score was also lower than Robust Wav2vec2.0. WavLM had the highest F1 scores for correct pronunciations and mispronunciations with 0.963 and 0.419 respectively. However,

the correlation of pronunciation scores was much lower than Robust Wav2vec2.0 and HuBERT with an average of 0.772.

To compare the three SSL models that were pre-trained only with English data, Robust Wav2vec2.0 had the best average performance for APA task, followed by HuBERT and WavLM. For MDD task, WavLM had the best performance followed by Robust Wav2vec2.0 and HuBERT. As these orders differ from the order in the amount of datasets used for pre-training, WavLM followed by Robust Wav2vec2.0 and HuBERT, the performance difference between these three SSL models may lie in their learning scheme.

Table 10 Experiment results for APA task with regard to using different backbone self-supervised learning model for the joint model (RMSE+GOP)

SSL model	Pronunciation Scores (PCC)			
	Accuracy	Fluency	Prosodic	Total
Robust (Baseline)	**0.751 ±0.004**	**0.815 ±0.002**	**0.810 ±0.002**	0.768 ±0.002
XLS-R	0.716 ±0.010	0.799 ±0.009	0.787 ±0.007	0.742 ±0.010
HuBERT	0.745 ±0.004	0.813 ±0.002	0.805 ±0.007	**0.769 ±0.005**
WavLM	0.732 ±0.008	0.806 ±0.003	0.794 ±0.002	0.755 ±0.008

Table 11 Experiment results for MDD task with regard to using different backbone self-supervised learning model for the joint model (RMSE+GOP)

SSL model	PER (%)	Correct Pronunciations			Mispronunciations		
		Prec.	Recall	F1	Prec.	Recall	F1
Robust	10.19	0.997	0.928	0.961	0.267	0.918	0.414

(Baseline)	±0.001	±0.000	±0.000	±0.000	±0.001	±0.002	±0.002
XLS-R	13.77 ±0.011	**0.998 ±0.000**	0.894 ±0.010	0.943 ±0.006	0.204 ±0.013	**0.946 ±0.011**	0.336 ±0.017
HuBERT	10.36 ±0.002	0.997 ±0.000	0.928 ±0.002	0.962 ±0.001	0.267 ±0.004	0.914 ±0.008	0.414 ±0.004
WavLM	**9.65 ±0.002**	0.997 ±0.000	**0.931 ±0.001**	**0.963 ±0.001**	**0.273 ±0.004**	0.903 ±0.004	**0.419 ±0.005**

4.4. Results Analysis

The experiment results are analyzed by looking at the accuracy of discrete scores for APA and discrete phones for MDD tasks, to see how the joint model showed performance improvement compared to single-task APA and MDD models. Joint-CAPT-SSL with RMSE loss and GOP features based on raw Robust Wav2vec2.0 is used for the analysis and is compared with its APA/MDD counterparts.

4.4.1. Analysis on model pronunciation assessment

Analysis on the performance of pronunciation assessment task is done with confusion matrices to see how the model prediction has improved for the joint model over single-task APA model. The confusion matrices were created using the trial that had the best average correlation performance among the three experiments, for each model. As the model predictions include decimal points, the predictions were

rounded to integer to form confusion matrix values.

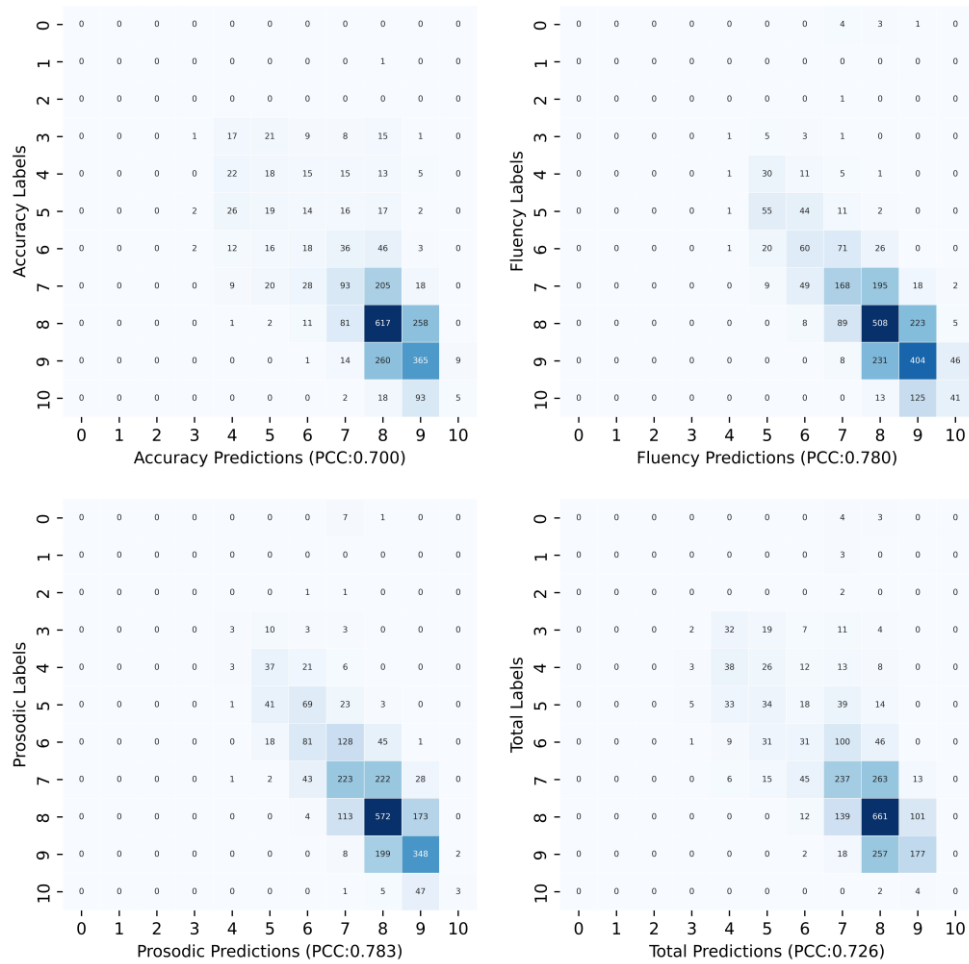

Figure 8 Confusion matrices of APA-SSL (RMSE+GOP)

Figure 8 and 9 show that the predicted scores of Joint-CAPT-SSL have overall better correlations than APA-SSL. The improvement is especially noticeable for the annotated label scores with low frequency (0-6, 10). This implies that the additional information of mispronunciations from joint learning assists the model to distinguish L2 pronunciations better, even for data with high class imbalance. The

confusion matrices of other models, APA-L1 and Joint-CAPT-L1, for RMSE+GOP architecture are provided in the Appendix.

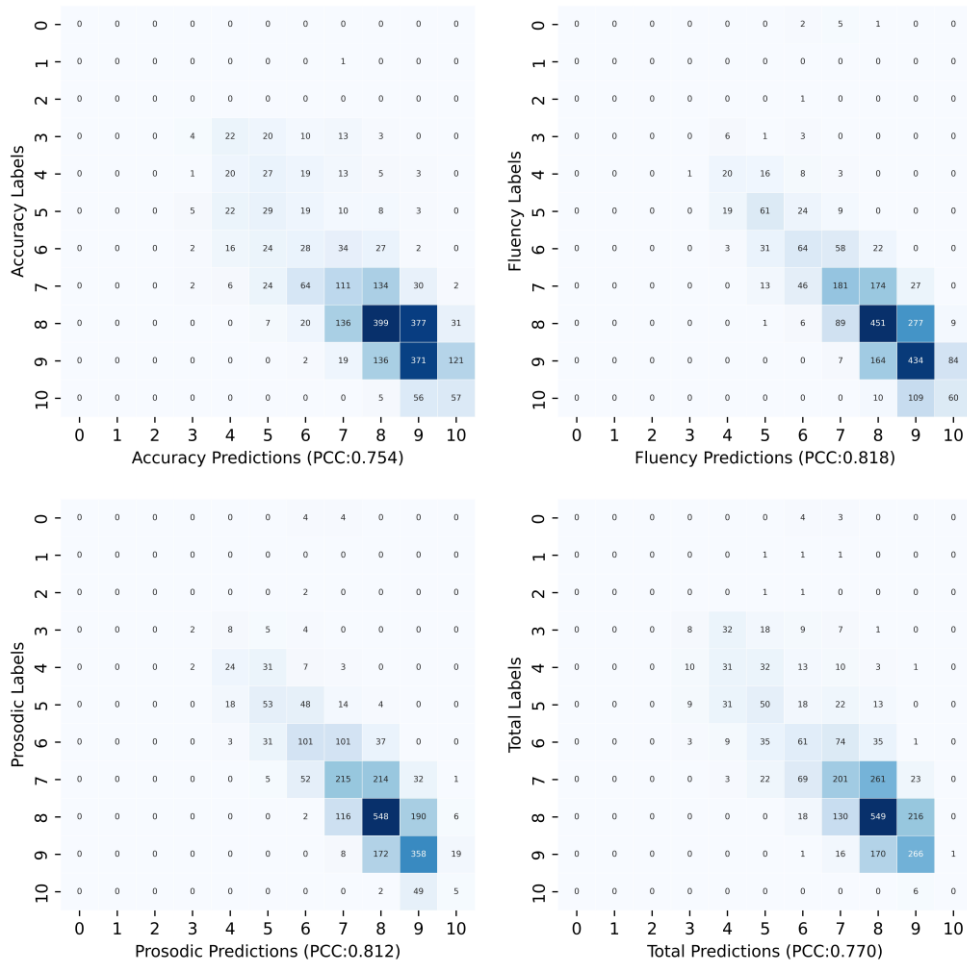

Figure 9 Confusion matrices of Joint-CAPT-SSL (RMSE+GOP)

4.4.2. Analysis on model mispronunciation detection and

diagnosis

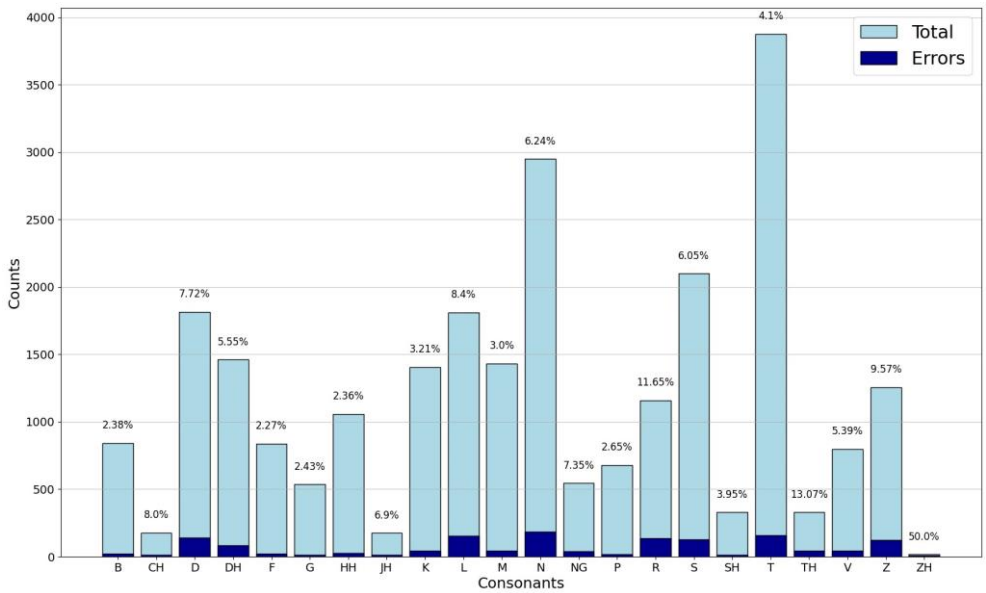

Figure 10 False Rejection rate of each consonant on correct pronunciations for Joint-CAPT-SSL (RMSE+GOP)

Analysis on the performance of mispronunciation detection and diagnosis task is done by looking at the error rate of each phones in the realized non-native speech labels to see the model's ability in distinguishing discrete phones. Error rates are counted with the trial of Joint-CAPT-SSL (RMSE+GOP) that showed the highest F1 scores for both correct pronunciations and mispronunciations. Thus, the False Negative rate is computed for all phones of correct pronunciations and mispronunciations, corresponding to False Rejection rate and False Acceptance rate, respectively. The phones are divided into consonants and vowels to be presented with its error rate above the bar graph as in Figure 10, 11, 12, and 13. The same figures for MDD-SSL (RMSE+GOP) are provided in the Appedix. <unk> is marked as ERR

for mispronunciations and is included with vowels.

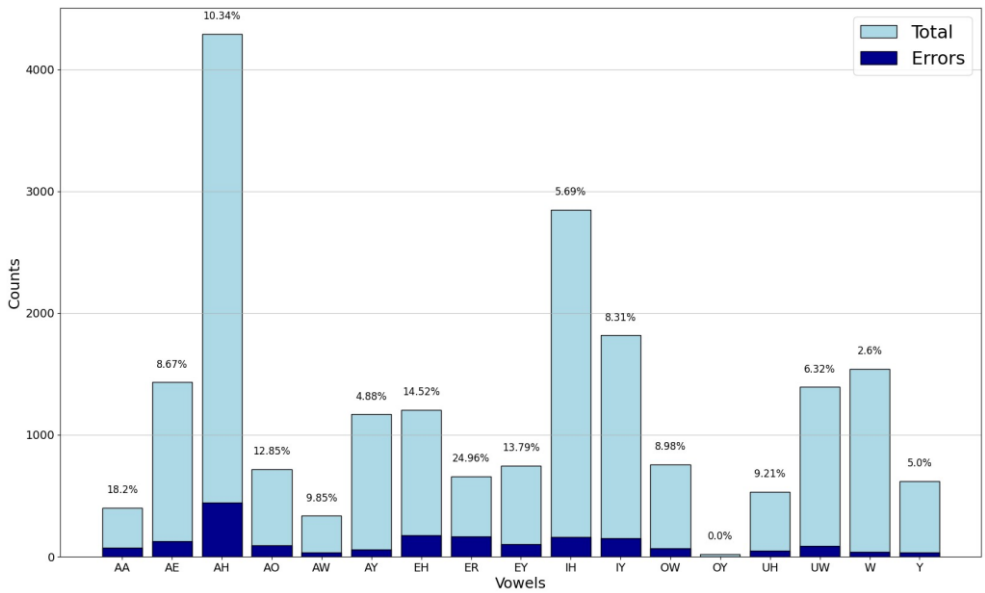

Figure 11 False Rejection rate of each vowel on correct pronunciations for Joint-CAPT-SSL (RMSE+GOP)

First for correct pronunciations, consonants showed lower False Rejection rate than vowels as shown in Figure 10 and 11, with an average rate of 7.83% and 9.66% normalized by phone occurrence. For consonants, ZH (50.00%), TH (13.07%), R (11.65%), Z (9.57%), L (8.40%), and CH (8.00%) had false rejection rate above the average error rate of consonants, meaning that models had more difficulty in distinguishing these consonants verbalized by non-native speakers. Interestingly, the frequency of occurrence was not the major cause of the result. Although ZH, TH, and CH are some of the labels with the lowest occurrence, R, Z, and L are labels with high frequency. This implies that the acoustic characteristics may be the underlying cause in the error rates. The model had difficulty in

recognizing fricatives (ZH, TH, Z), liquids (R, L) and affricates (CH). For vowels, the model showed higher False Rejection rate for ER (24.96%), AA (18.20%), EH (14.52%), EY (13.79%), AO (12.85%), AH (10.34%), and AW (9.75%) than the average error rate for vowels. Again, the frequency of occurrence for each vowel did not contribute in the error rates, with ER, AA, EY, AO, and AW showing low occurrence but EH and AH showing high occurrence. The model overall had difficulty in distinguishing mid, open vowels.

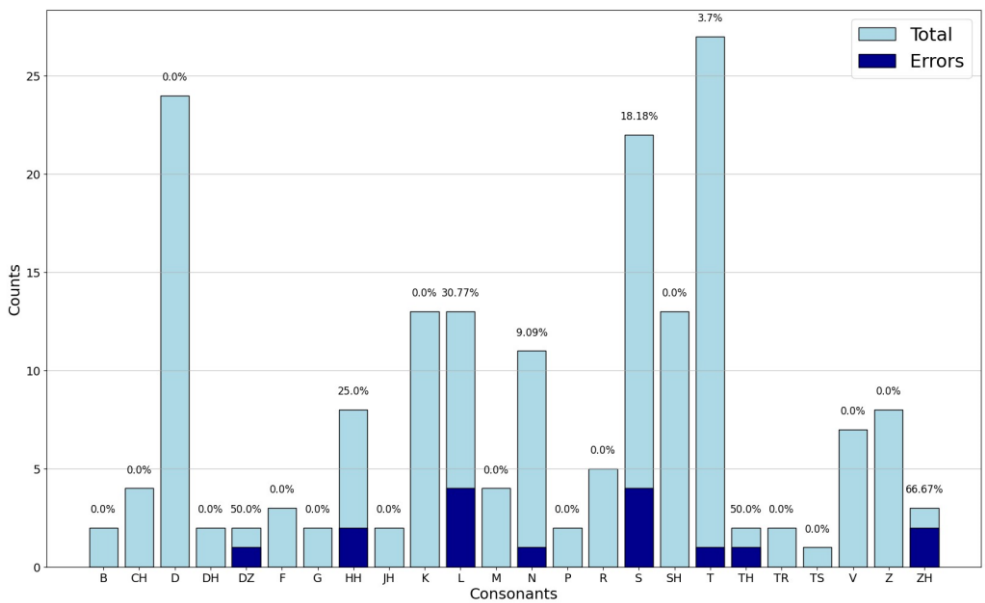

Figure 12 False Acceptance rate of each consonant on mispronunciations for Joint-CAPT-SSL (RMSE+GOP)

Second, mispronunciations also showed a similar tendency, with consonants showing lower False Acceptance rates than vowels as shown in Figure 12 and 13, with an average rate of 10.56% and 12.47% normalized by phone occurrence. For consonants, ZH (66.67%), TH (50.00%), DZ (50.00%), L (30.77%),

HH (25.00%), and S (18.18%) showed False Acceptance rate above the average error rate of consonants, meaning that models predicted these annotated mispronunciations to correct canonical pronunciations. As with correct pronunciations, the frequency of occurrence did not influence False Acceptance rates, with L and S having high occurrences. The model again showed weak performances in recognizing fricatives (ZH, TH, DZ, HH, S) and liquids (L). However, with the total mispronounced consonant annotations summing up to only 182, more mispronunciation annotations would be necessary to confirm this. For vowels, the model showed higher False Acceptance rate for Y (50.00%), OW (30.30%), UW (28.57%), UH (25.00%), AY (18.15%), AO (18.18%), AA (17.86%), and ER (14.81%) than the average for vowels, with the occurrence frequency mixed between low (Y, UW, UH, AY, AO) and high (OW, AA, ER). Similar to correct pronunciations, the model was weak at mid, open vowels but with only a total of 743 mispronounced vowel annotations, more mispronunciation annotated datasets would be necessary to verify this along with consonant mispronunciations.

Overall, phone error rates were not significantly high, with 7.74% for correct pronunciations and 9.99% for mispronunciations. These are lower than the phone error rates of MDD-SSL (RMSE+GOP) which showed an average error rate of 7.83% for correct pronunciations and 11.98% for mispronunciations. The results show that Joint-CAPT-SSL (RMSE+GOP) showed improvements in distinguishing discrete phones, especially for mispronunciations, which led to higher Precision for mispronunciations as in Table 7. However, due to the extreme imbalance between two ground-truth annotations, the overall Precision for mispronunciations remains low as the absolute value of False Rejection (the dark blue part for correct

pronunciation) is much bigger than True Rejection (the light blue part for mispronunciations) that together consist the formula of Precision for mispronunciations.

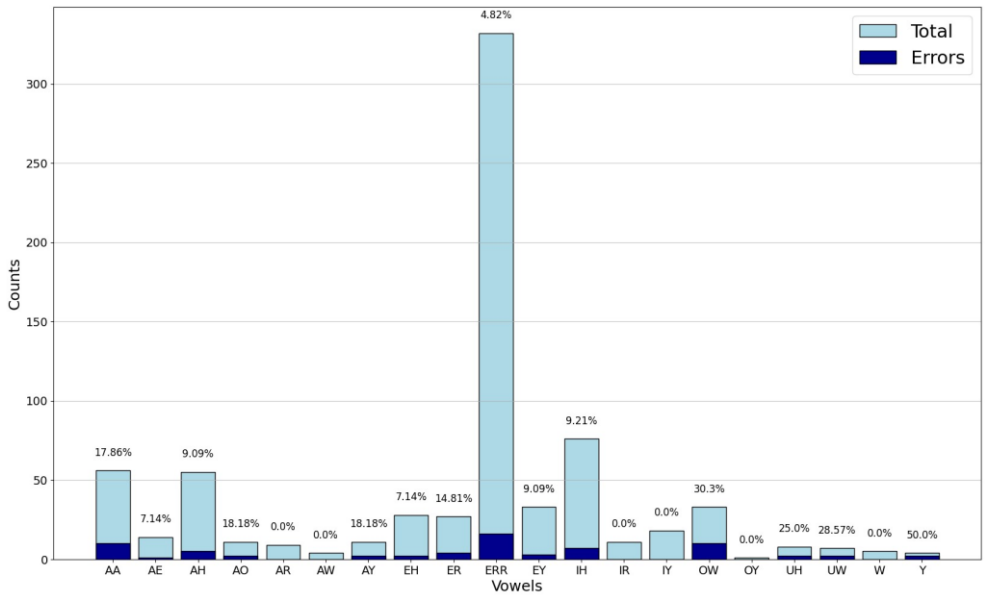

Figure 13 False Acceptance rate of each vowel on mispronunciations for Joint-CAPT-SSL (RMSE+GOP)

Chapter 5. Discussion

5.1. Correlation analysis on human assessments and model assessments

To explore how the joint model leverages the correlation between pronunciation scores and mispronunciations to improve APA and MDD performances, correlation analysis was conducted using Pearson correlation coefficients for both human evaluators and predictions of the proposed Joint-CAPT-SSL (RMSE+GOP). The test set of Speechocean762 was used for analysis. The results are plotted using linear regression in Figure 14. For both plots, accuracy, fluency, prosodic, and total score all showed a correlation with mispronunciations and were statistically significant ($p<.001$).

Specifically, for human evaluators, the total score had the highest negative correlation with mispronunciations ($r=-0.656$), followed by accuracy ($r=-0.624$), fluency ($r=-0.606$), and prosodic ($r=-0.593$). This suggests that the human assessors of Speechocean762 were influenced by phone errors when grading the scores for all aspects, which complies with the findings of previous linguistic findings.

Interestingly, model predictions also showed a similar pattern where accuracy ($r=-0.535$) and total score ($r=-0.535$) had the highest negative correlation with mispronunciation, followed by fluency ($r=-0.520$) and prosodic ($r=-0.519$). This also corresponds to the performance results of the model on Table 4, where accuracy and total score gained the most performance increase compared to APA-SSL.

Furthermore, the strength of the correlation between the predicted

pronunciation scores and mispronunciations increased as the APA performance improved for the joint model. To compare the correlation results between joint model architectures, Joint-CAPT-L1 (CE), Joint-CAPT-SSL (RMSE+GOP), the model predictions became more negatively correlated for fluency (-0.461, -0.520), prosodic (-0.476, -0.519), and total (-0.534, -0.535) as the PCC increased for the latter model. This proves that the joint model has learned the behavior of human evaluators in non-native assessments while improving the performance. The linear regression plot for Joint-CAPT-L1 (CE) is provided in the Appendix.

Altogether, the statistical analysis provides evidence that the integrated model leveraged the correlation between APA and MDD tasks to gain performance improvement.

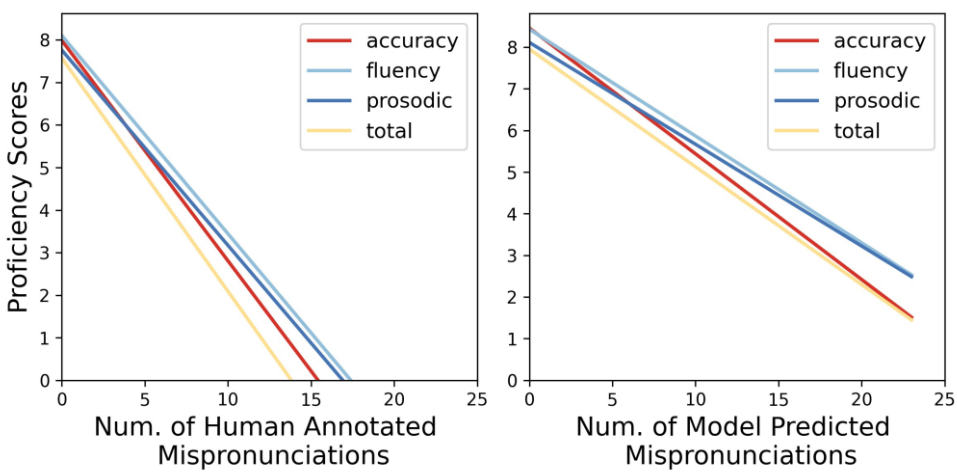

Figure 14 Correlation between pronunciation scores of four aspects and the number of mispronunciations predicted by Joint-CAPT-SSL (RMSE+GOP)

5.2. Analysis on multi-task learning loss weight

The effect of giving different weights on pronunciation assessment task is presented in Table 12 and 13 by controlling the size of α on the loss L_{APA} for the Joint-CAPT-SSL (RMSE+GOP). The results show that decreasing the weight on L_{APA} (0.1, 0.25, 0.5) helps the model achieve better results on fluency and prosodic correlation, with better MDD performances as well. Specifically, fluency and prosodic scores achieved the highest correlation with 0.818 and 0.812 when α was set to 0.1. F1 score for correct pronunciations was the highest when α = 0.05, 0.1, 0.25, 0.5, and F1 score for mispronunciations was the highest when α = 0.25. However as the weight on pronunciation assessment gets smaller from 0.25 to 0.1, there exists a steep decrease in the performance on accuracy and total scores.

Increasing the weight on L_{APA} (0.25, 0.5, 1.0) results in better accuracy and total scores. Accuracy score was the highest when α = 0.25 and total score had the best correlation with human experts when the weight was α = 0.25, 0.5. However, similar to how smaller weights caused a fall in accuracy and total scores, higher weight leads to a decrease in fluency, prosodic scores and F1 scores, as in the comparison between α = 0.25 and α = 1.0

Yet, a simple decrease or increase in weights do not necessarily guarantee higher fluency, prosodic scores, and F1 score or higher accuracy and total scores. The performance for fluency, prosodic scores and F1 score for mispronunciations decreased when the weights were too low (0.05), and the performance for accuracy and total scores decreased as the weight became too big. (1.5, 2.0) Overall, α = 0.25 achieved the most decent performance for both APA and MDD.

This is in line with the results from the correlation analysis. The correlation

between mispronunciations and accuracy (r=-0.624) and total (r=-0.656) scores were relatively higher than fluency (r=-0.606) and prosodic (r=-0.593) for human experts. This is reflected in the results with bigger weights on pronunciation assessment task leading to better APA results for accuracy and total scores. The results on different multi-task learning loss weights imply that the joint model not only learned to leverage the relationship between mispronunciations and pronunciation scores to gain performance increases for APA and MDD tasks compared to respective single-task models as shown in the correlation analysis, but also learned the different degree of the correlation between the individual scores and mispronunciations while training.

Table 12 Experiment results for APA task with regard to different multi-task learning loss weight for Joint-CAPT-SSL (RMSE+GOP)

MTL loss weight	Pronunciation Scores (PCC)			
	Accuracy	Fluency	Prosodic	Total
0.05	0.730 ±0.001	0.797 ±0.002	0.792 ±0.003	0.747 ±0.002
0.1	0.744 ±0.004	**0.818 ±0.003**	**0.812 ±0.004**	0.762 ±0.005
0.25 (Baseline)	**0.751 ±0.004**	0.815 ±0.002	0.810 ±0.002	**0.768 ±0.002**
0.5	0.750 ±0.003	0.816 ±0.002	0.811 ±0.003	**0.768** ±0.003
1.0	0.750 ±0.002	0.809 ±0.002	0.806 ±0.003	0.766 ±0.005
1.5	0.747 ±0.006	0.809 ±0.002	0.807 ±0.003	0.767 ±0.003
2.0	0.742 ±0.006	0.809 ±0.004	0.807 ±0.002	0.761 ±0.004

Table 13 Experiment results for MDD task with regard to different multi-task learning loss weight for Joint-CAPT-SSL (RMSE+GOP)

MTL loss weight	PER (%)	Correct Pronunciations			Mispronunciations		
		Precision	Recall	F1	Precision	Recall	F1
0.05	10.30 ±0.000	0.997 ±0.000	0.927 ±0.001	**0.961 ±0.000**	0.265 ±0.002	0.916 ±0.001	0.411 ±0.003
0.1	**10.18 ±0.001**	0.997 ±0.000	**0.928 ±0.001**	**0.961 ±0.001**	**0.267 ±0.004**	0.917 ±0.002	0.413 ±0.004
0.25 (Baseline)	10.19 ±0.001	0.997 ±0.000	**0.928 ±0.000**	**0.961 ±0.000**	**0.267 ±0.001**	0.918 ±0.002	**0.414 ±0.002**
0.5	10.25 ±0.001	**0.998 ±0.000**	0.927 ±0.001	**0.961 ±0.001**	0.266 ±0.003	0.925 ±0.004	0.413 ±0.005
1.0	10.42 ±0.001	**0.998 ±0.000**	0.926 ±0.001	0.960 ±0.005	0.262 ±0.003	0.921 ±0.005	0.408 ±0.003
1.5	10.77 ±0.001	**0.998 ±0.000**	0.922 ±0.002	0.959 ±0.001	0.255 ±0.006	0.925 ±0.005	0.400 ±0.008
2.0	11.00 ±0.001	**0.998 ±0.000**	0.920 ±0.001	0.957 ±0.001	0.251 ±0.002	**0.931 ±0.004**	0.395 ±0.002

Chapter 6. Conclusion

This study presents a novel architecture that jointly trains automatic pronunciation assessment and mispronunciation detection and diagnosis with a multi-task learning perspective, motivated by the high linguistic correlation between pronunciation scores and mispronunciations. The proposed joint model shows significant performance improvement over single-task APA and MDD on Speechocean762, by learning to better distinguish pronunciation scores with low distribution and to better recognize mispronunciations. This proves that the correlation between two tasks can benefit each other, which is supported by the correlation analysis and the multi-task learning loss weight analysis. The proposed model not only conforms to the linguistic mechanism of non-native speech assessment, but shows its usefulness in practical assessment scenarios where learners are graded in various aspects with a single utterance.

References

Algabri, Mohammed, Hassan Mathkour, Mansour Alsulaiman, and Mohamed A. Bencherif. 2022. "Mispronunciation Detection and Diagnosis with Articulatory-Level Feedback Generation for Non-Native Arabic Speech." *Mathematics* 10(15):2727. doi: 10.3390/math10152727.

Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale."

Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale." Pp. 2278–82 in *Interspeech 2022*. ISCA.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations."

Cai, Xingyu, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. 2021. "Speech Emotion Recognition with Multi-Task Learning." Pp. 4508–12 in *Interspeech 2021*. ISCA.

Carnegie Mellon University. "The CMU Pronouncing Dictionary." 2000. Retrieved July 1, 2023 (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).

Chao, Fu-An, Tien-Hong Lo, Tzu-I. Wu, Yao-Ting Sung, and Berlin Chen. 2022. "3M: An Effective Multi-View, Multi-Granularity, and Multi-Aspect Modeling Approach to English Pronunciation Assessment." Pp. 575–82 in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Chiang Mai, Thailand: IEEE.

Chen, Lei, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. "End-to-End Neural Network Based Automated Speech Scoring." Pp. 6234–38 in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, Nancy F., Darren Wee, Rong Tong, Bin Ma, and Haizhou Li. 2016. "Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: Analysis on ICALL." *Speech Communication* 84:46–56. doi: 10.1016/j.specom.2016.07.005.

Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen,

Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing." *IEEE Journal of Selected Topics in Signal Processing* 16(6):1505–18. doi: 10.1109/JSTSP.2022.3188113.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. "Unsupervised Cross-Lingual Representation Learning for Speech Recognition." Pp. 2426–30 in *Interspeech 2021*. ISCA.

van Dalen, Rogier C., Kate M. Knill, and Mark J. F. Gales. 2015. "Automatically Grading Learners' English Using a Gaussian Process."

Eskenazi, Maxine. 2009. "An Overview of Spoken Language Technology for Education." *Speech Communication* 51(10):832–44. doi: 10.1016/j.specom.2009.04.005.

Fan, Zhixing, Jing Li, Aishan Wumaier, Zaokere Kadeer, and Abdujelil Abdurahman. 2023. "A Multifaceted Approach to Oral Assessment Based on the Conformer Architecture." *IEEE Access* 11:28318–29. doi: 10.1109/ACCESS.2023.3255986.

Feng, Yiqing, Guanyu Fu, Qingcai Chen, and Kai Chen. 2020. "SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis." Pp. 3492–96 in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. 1993. "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1." *NASA STI/Recon Technical Report N* 93:27403.

Gong, Yuan, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. "Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment." Pp. 7262–66 in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. "Conformer: Convolution-Augmented Transformer for Speech Recognition." Pp. 5036–40 in *Interspeech 2020*. ISCA.

Guo, Shen, Zaokere Kadeer, Aishan Wumaier, Liejun Wang, and Cong Fan. 2023. "Multi-Feature and Multi-Modal Mispronunciation Detection and Diagnosis Method Based on the Squeezeformer Encoder." *IEEE Access* 1–1. doi: 10.1109/ACCESS.2023.3278837.

Harrison, Alissa M., Wai-kit Lo, Xiao-jun Qian, and Helen Meng. 2009. "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training."

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29(6):82–97. doi: 10.1109/MSP.2012.2205597.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29:3451–60. doi: 10.1109/TASLP.2021.3122291.

Hsu, Wei-Ning, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. "Robust Wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training."

Hu, Wenping, Yao Qian, and Frank K. Soong. 2015. "An Improved DNN-Based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech."

Kim, Eesung, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. "Automatic Pronunciation Assessment Using Self-Supervised Speech Representation Learning." Pp. 1411–15 in *Interspeech 2022*. ISCA.

Kim, Sehoon, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. 2022. "Squeezeformer: An Efficient Transformer for Automatic Speech Recognition."

Kyriakopoulos, Konstantinos, Kate Knill, and Mark Gales. 2018. "A Deep Learning Approach to Assessing Non-Native Pronunciation of English Using Phone Distances." Pp. 1626–30 in *Interspeech 2018*. ISCA.

Leung, Wai-Kim, Xunying Liu, and Helen Meng. 2019. "CNN-RNN-CTC Based End-to-End Mispronunciation Detection and Diagnosis." Pp. 8132–36 in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Li, Kun, Xiaojun Qian, and Helen Meng. 2017. "Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language*

Processing 25(1):193–207. doi: 10.1109/TASLP.2016.2621675.

Lin, Binghuai, and Liyuan Wang. 2021a. *A Noise Robust Method for Word-Level Pronunciation Assessment*.

Lin, Binghuai, and Liyuan Wang. 2021b. "Deep Feature Transfer Learning for Automatic Pronunciation Assessment." Pp. 4438–42 in *Interspeech 2021*. ISCA.

Lin, Binghuai, and Liyuan Wang. 2022. "Phoneme Mispronunciation Detection By Jointly Learning To Align." Pp. 6822–26 in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lin, Binghuai, and Liyuan Wang. 2023. "Exploiting Information From Native Data for Non-Native Automatic Pronunciation Assessment." Pp. 708–14 in *2022 IEEE Spoken Language Technology Workshop (SLT)*.

Lin, Binghuai, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang. 2020. "Automatic Scoring at Multi-Granularity for L2 Pronunciation." Pp. 3022–26 in *Interspeech 2020*. ISCA.

Litman, Diane, Helmer Strik, and Gad S. Lim. 2018. "Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities." *Language Assessment Quarterly* 15(3):294–309. doi: 10.1080/15434303.2018.1472265.

Lo, Tien-Hong, Shi-Yan Weng, Hsiu-Jui Chang, and Berlin Chen. 2020. "An Effective End-to-End Modeling Approach for Mispronunciation Detection." Pp. 3027–31 in *Interspeech 2020*. ISCA.

Metallinou, Angeliki, and Jian Cheng. 2014. "Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners." Pp. 1468–72 in *Interspeech 2014*. ISCA.

Minematsu, Nobuaki, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu, Seiichi Nakagawa, Masatake Dantsuji, and Shozo Makino. n.d. "Development of English Speech Database Read by Japanese to Support CALL Research."

Munro, Murray J., and Tracey M. Derwing. 1995. "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners." *Language Learning* 45(1):73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x.

Naijo, Satsuki, Akinori Ito, and Takashi Nose. 2021. "Improvement of Automatic English Pronunciation Assessment with Small Number of Utterances Using Sentence Speakability." Pp. 4473–77 in *Interspeech 2021*. ISCA.

O'Brien, Mary Grantham. 2014. "L2 Learners' Assessments of Accentedness, Fluency, and Comprehensibility of Native and Nonnative German Speech." *Language Learning* 64(4):715–48. doi: 10.1111/lang.12082.

Peng, Linkai, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. "A Study on Fine-Tuning Wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis." Pp. 4448–52 in *Interspeech 2021*. ISCA.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukasˇ Burget, Ondˇrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlıcˇek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. n.d. "The Kaldi Speech Recognition Toolkit."

Rogerson-Revell, Pamela M. 2021. "Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions." *RELC Journal* 52(1):189–205. doi: 10.1177/0033688220977406.

Ryu, Hyuksu, and Minhwa Chung. 2017. "Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features." Pp. 65–70 in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*. ISCA.

Ryu, Hyuksu, Hyejin Hong, Sunhee Kim, and Minhwa Chung. 2016. "Automatic Pronunciation Assessment of Korean Spoken by L2 Learners Using Best Feature Set Selection." Pp. 1–6 in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Jeju, South Korea: IEEE.

Shi, Jiatong, Nan Huo, and Qin Jin. 2020. "Context-Aware Goodness of Pronunciation for Computer-Assisted Pronunciation Training." Pp. 3057–61 in *Interspeech 2020*. ISCA.

Sudhakara, Sweekar, Manoj Kumar Ramanathi, Chiranjeevi Yarra, Anurag Das, and Prasanta Kumar Ghosh. 2019. "Noise Robust Goodness of Pronunciation Measures Using Teacher's Utterance." Pp. 69–73 in *SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*. ISCA.

Sudhakara, Sweekar, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2019. "An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities." Pp. 954–58 in *Interspeech 2019*. ISCA.

Tong, Rong, Boon Pang Lim, Nancy F. Chen, Bin Ma, and Haizhou Li. 2014. "Subspace Gaussian Mixture Model for Computer-Assisted Language Learning." Pp. 5347–51 in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tsubota, Yasushi, Masatake Dantsuji, and Tatsuya Kawahara. 2004. "*An English Pronunciation Learning System for Japanese Students Based on Diagnosis of Critical Pronunciation Errors*." *ReCALL* 16(1):173–88. doi: 10.1017/S0958344004001314.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." in *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Wadud, Md Anwar Hussen, Mohammed Alatiyyah, and M. F. Mridha. 2023. "Non-Autoregressive End-to-End Neural Modeling for Automatic Pronunciation Error Detection." *Applied Sciences* 13(1):109. doi: 10.3390/app13010109.

Wang, Hsin-Wei, Bi-Cheng Yan, Hsuan-Sheng Chiu, Yung-Chang Hsu, and Berlin Chen. 2022. "Exploring Non-Autoregressive End-to-End Neural Modeling for English Mispronunciation Detection and Diagnosis." Pp. 6817–21 in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Witt, S. M., and S. J. Young. 2000. "Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning." *Speech Communication* 30(2–3):95–108. doi: 10.1016/S0167-6393(99)00044-8.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. "Transformers: State-of-the-Art Natural Language Processing." Pp. 38–45 in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics.

Xu, Xiaoshuo, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. "Explore Wav2vec 2.0 for Mispronunciation Detection." Pp. 4428–32 in *Interspeech 2021*. ISCA.

Yang, Seung Hee, and Minhwa Chung. 2017. "Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency." Pp. 53–58 in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*. ISCA.

Yeo, Eun Jung, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. 2023. "Automatic Severity Classification of Dysarthric Speech by Using Self-Supervised Model with Multi-Task Learning." Pp. 1–5 in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zechner, Klaus, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. "Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English." *Speech Communication* 51(10):883–95. doi: 10.1016/j.specom.2009.04.009.

Zhang, Daniel, Ashwinkumar Ganesan, Sarah Campbell, and Daniel Korzekwa. 2022. "L2-GEN: A Neural Phoneme Paraphrasing Approach to L2 Speech Synthesis for Mispronunciation Diagnosis." Pp. 4317–21 in *Interspeech 2022*. ISCA.

Zhang, Junbo, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. "Speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment." Pp. 3710–14 in *Interspeech 2021*. ISCA.

Zhang, Long, Ziping Zhao, Chunmei Ma, Linlin Shan, Huazhi Sun, Lifen Jiang, Shiwen Deng, and Chang Gao. 2020. "End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC/Attention Architecture." *Sensors (14248220)* 20(7):1809. doi: 10.3390/s20071809.

Zhao, Guanlong, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. "L2-ARCTIC: A Non-Native English Speech Corpus." Pp. 2783–87 in *Interspeech 2018*. ISCA.

Zheng, Nianzu, Liqun Deng, Wenyong Huang, Yu Ting Yeung, Baohua Xu, Yuanyuan Guo, Yasheng Wang, Xiao Chen, Xin Jiang, and Qun Liu. 2022. "CoCA-MDD: A Coupled Cross-Attention Based Framework for Streaming Mispronunciation Detection and Diagnosis." Pp. 4352–56 in *Interspeech 2022*. ISCA.

Appendix

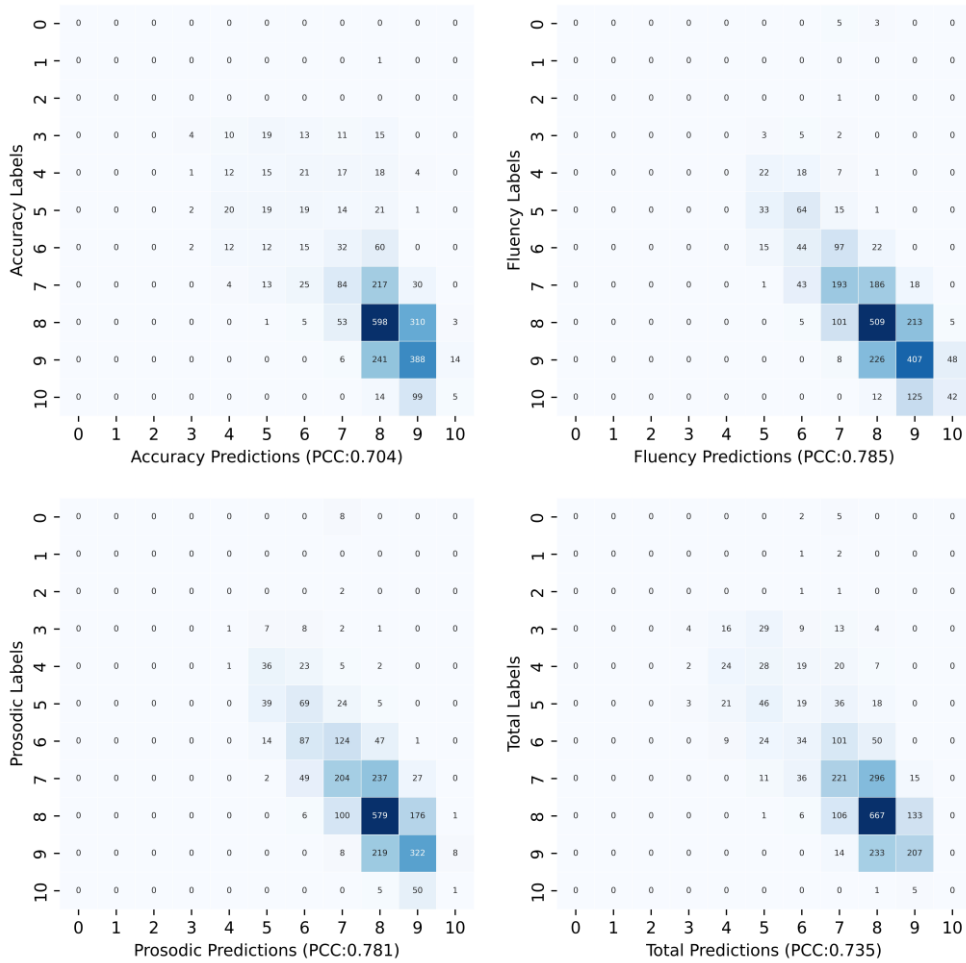

Figure 15 Confusion matrices of APA-L1 (RMSE+GOP)

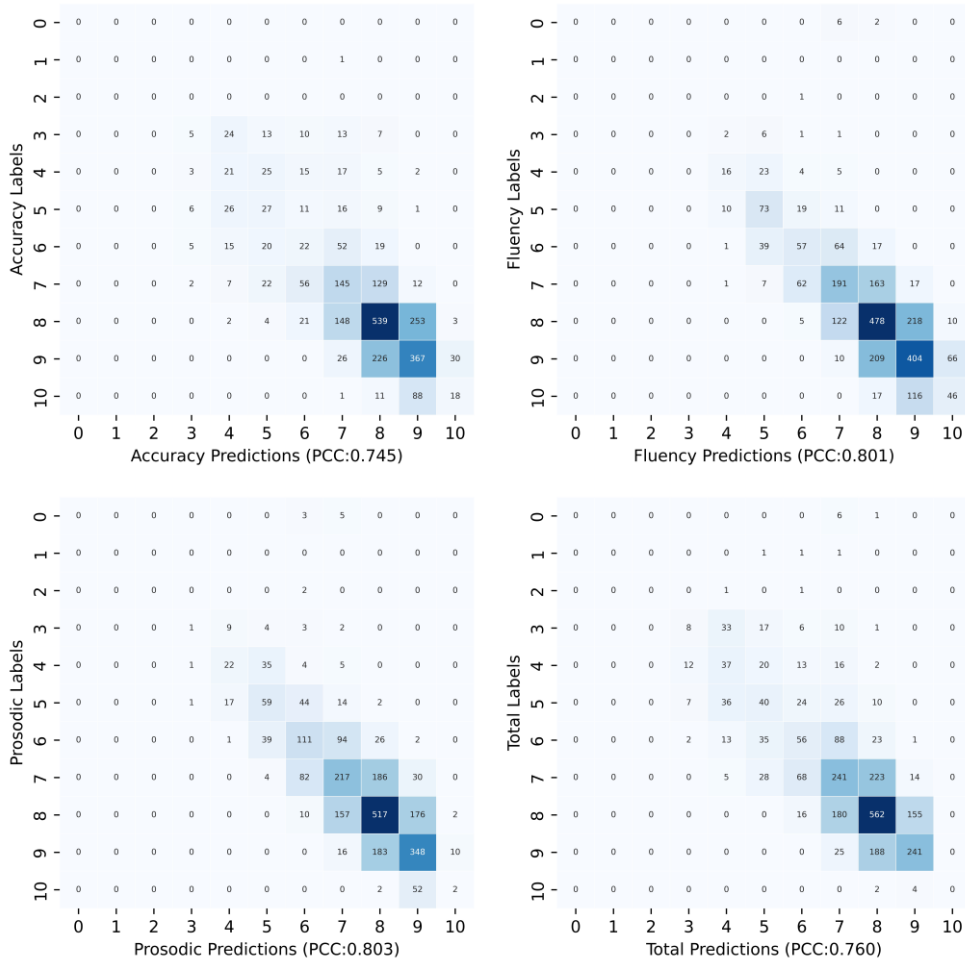

Figure 16 Confusion matrices of Joint-CAPT-L1 (RMSE+GOP)

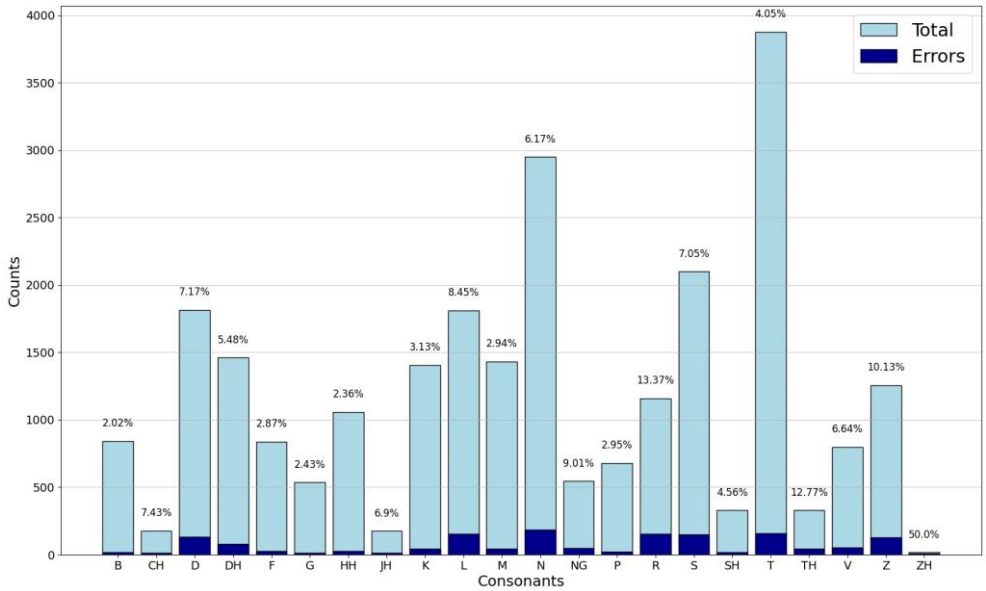

Figure 17 False Rejection rate of each consonant on correct pronunciations for MDD-SSL (RMSE+GOP)

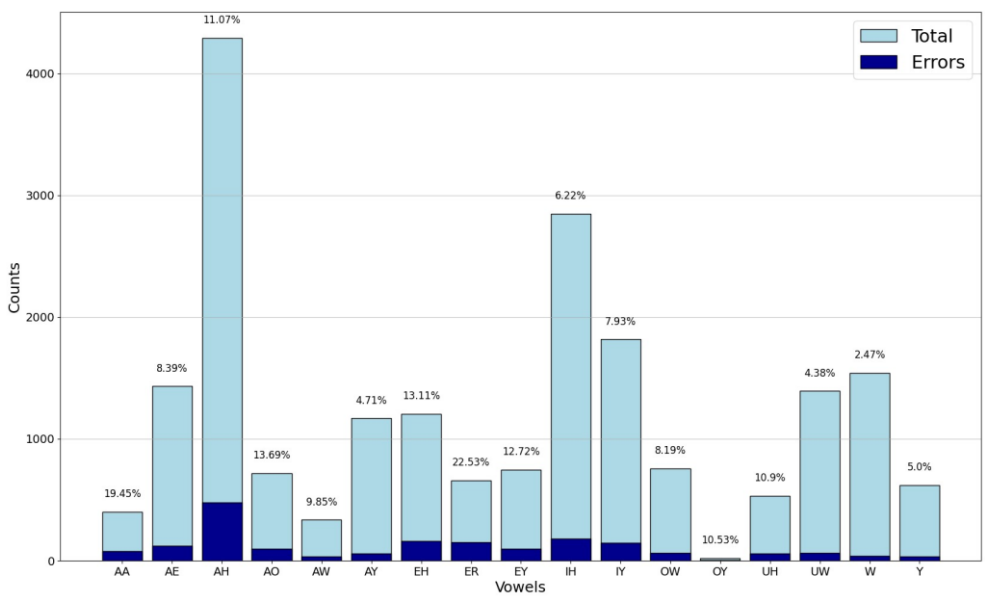

Figure 18 False Rejection rate of each vowel on correct pronunciations for MDD-SSL

(RMSE+GOP)

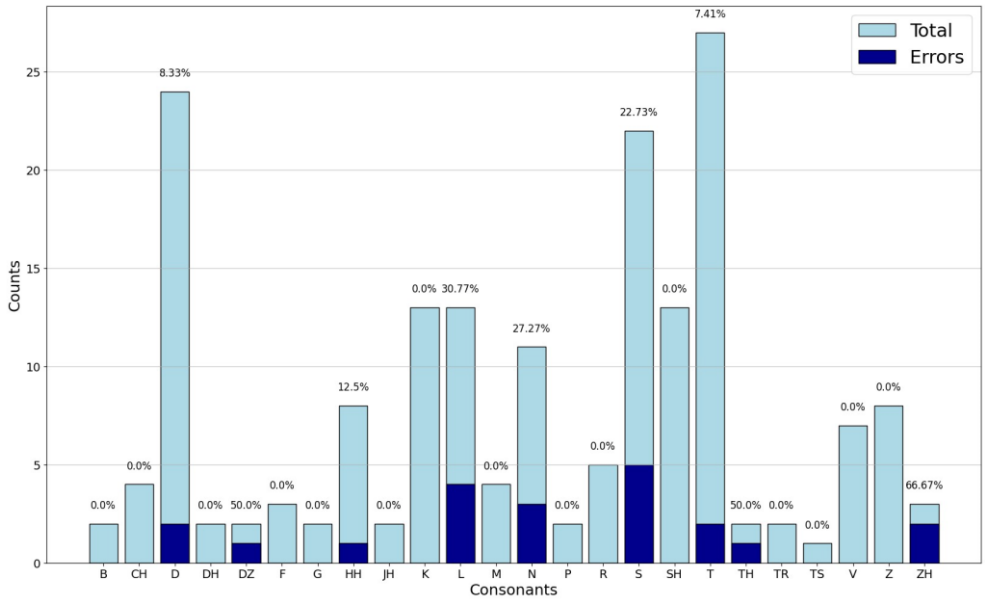

Figure 19 False Acceptance rate of each consonant on mispronunciations for MDD-SSL

(RMSE+GOP)

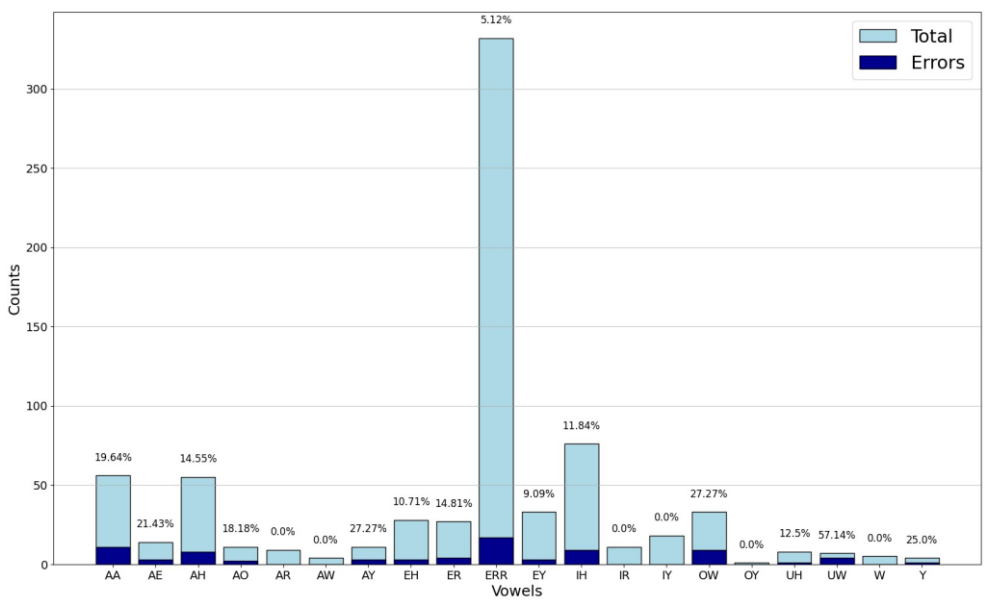

Figure 20 False Acceptance rate of each vowel on mispronunciations for MDD-SSL

(RMSE+GOP)

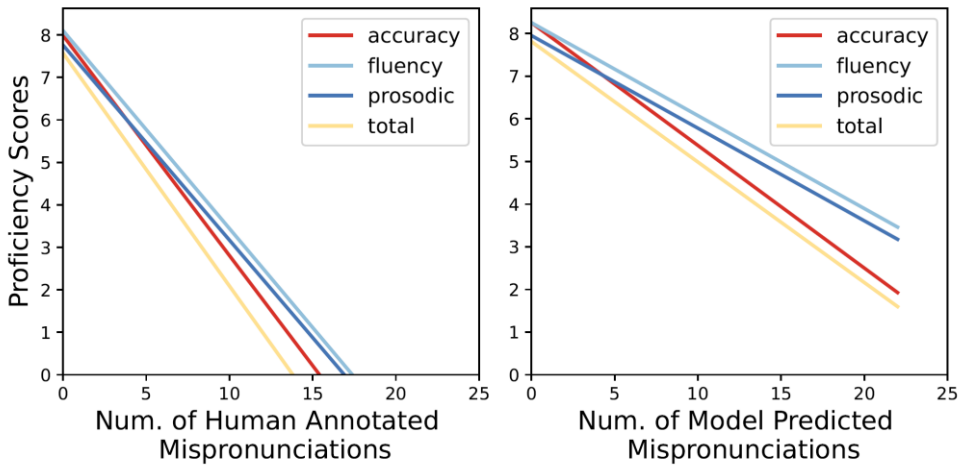

Figure 21 Correlation between pronunciation scores of four aspects and the number of mispronunciations predicted by Joint-CAPT-L1 (CE)

국문 초록

실증 연구에 의하면 비원어민 발음 평가에 있어 전문 평가자가 채점하는 발음 점수와 음소 오류 사이의 상관관계는 매우 높다. 그러나 기존의 컴퓨터기반발음훈련 (Computer-assisted Pronunciation Training; CAPT) 시스템은 자동발음평가 (Automatic Pronunciation Assessment; APA) 과제 및 발음오류검출 (Mispronunciation Detection and Diagnosis; MDD) 과제를 독립적인 과제로 취급하며 각 모델의 성능을 개별적으로 향상시키는 것에만 초점을 두었다. 본 연구에서는 두 과제 사이의 높은 상관관계에 주목, 다중작업학습 기법을 활용하여 자동발음평가와 발음오류검출 과제를 동시에 훈련하는 새로운 아키텍처를 제안한다. 구체적으로는 APA 과제를 위해 교차 엔트로피 손실함수 및 RMSE 손실함수를 실험하며, MDD 손실함수는 CTC 손실함수로 고정된다. 근간 음향 모델은 사전훈련된 자기지도학습기반 모델로 하며, 이때 더욱 풍부한 음향 정보를 위해 다중작업학습을 거치기 전에 부수적으로 음소인식에 대하여 미세조정되기도 한다. 음향 모델과 함께 발음적합점수(Goodness-of-Pronunciation; GOP)가 추가적인 입력으로 사용된다.

실험 결과, 통합 모델이 단일 자동발음평가 및 발음오류검출 모델보다 매우 높은 성능을 보였다. 구체적으로는 Speechocean762 데이터셋에서 자동발음평가 과제에 사용된 네 항목의 점수들의 평균 피어슨상관계수가 0.041 증가하였으며, 발음오류검출 과제에 대해 F1 점수가 0.003 증가하였다. 통합 모델에 대해 시도된 아키텍처 중에서는, Robust Wav2vec2.0 음향모델과 발음적합점수를 활용하여 RMSE/CTC 손실함수로 훈련한 모델의 성능이 가장 좋았다. 모델을 분석한 결과, 통합 모델이 개별 모델에 비해 분포가 낮은 점수 및 발음오류를 더 정확하게 구분하였음을 확인할 수 있었다.

흥미롭게도 통합 모델에 있어 각 하위 과제들의 성능 향상 정도는 각 발음 점수와 발음 오류 레이블 사이의 상관계수 크기에 비례하였다. 또 통합 모델의 성능이 개선될수록 모델의 예측 발음점수, 그리고 모델의 예측 발음오류에 대한 상관성이 높아졌다. 본 연구 결과는 통합 모델이 발음 점수 및 음소 오류 사이의 언어학적 상관성을 활용하여 자동발음평가 및 발음오류검출 과제의 성능을 향상시켰으며, 그 결과 통합 모델이 전문 평가자들의 실제 비원어민 평가와 비슷한 양상을 띤다는 것을 보여준다.