論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Towards Self-Supervised Learning based Acoustic Modeling for Non-Native Mispronunciation Verification
著者(和文)	YANGLongfei
Author(English)	Longfei Yang
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第284号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:篠﨑 隆宏,奥村 学,中山 実,船越 孝太郎,長谷川 晶一
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第284号, Conferred date:2025/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Institute of Science Tokyo

Department of Information and Communication Engineering

Towards Self-Supervised Learning based Acoustic Modeling for Non-Native Mispronunciation Verification

BY

Longfei Yang

A Doctoral Thesis Submitted to the Faculty of

Department of Information and Communications Engineering

In Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Engineering

December 2024

Abstract

The growing demand for learning a second language (L2) in today's globalized and socially integrated world has brought significant attention to computer-aided language learning (CALL) systems. Compared to traditional one-on-one communicative approaches between teachers and students in classroom, CALL systems offer greater flexibility while saving resources such as teaching staff, administrative personnel, and classroom space. A vital component of these systems is computer-aided pronunciation training (CAPT), which functions as a virtual instructor. CAPT processes and analyzes learners' speech, assesses pronunciation quality, and provides targeted feedback for improvement, commonly referred to as pronunciation assessment and mispronunciation verification.

Effective CAPT systems should not only detect pronunciation errors in learners' non-native utterances but also diagnose the type and location of these errors. Furthermore, they should provide actionable feedback to help learners correct their mispronunciations. For example, if a student pronounces the vowel "u" incorrectly, spreading their lips instead of rounding them, the system should not only identify this error but also offer guidance, such as: "Try not to spread your lips when pronouncing the rounded sound /u/." Research shows that incorporating information about the place and manner of articulation in feedback gives learners a clear understanding of how to adjust their articulators for correct pronunciation. This approach mirrors the instructional quality of a professional language teacher.

Mispronunciation verification is central to the CAPT system, with most current systems leveraging state-of-the-art speech technologies to establish acoustic models for this task. While developing acoustic models using non-native speech data is conceptually straightforward, the practical challenge lies in the scarcity of large, annotated datasets. Collecting and labeling non-native speech requires significant time and manual effort, creating a data sparsity issue that hampers the performance of supervised learning approaches for mispronunciation verification.

To address this challenge, two main research directions have been explored. The first focuses on techniques to maximize the utility of limited non-native speech data, while the second relies on transfer learning. Transfer learning typically involves pre-training a model on a large dataset for a general task, such as speech recognition, and then fine-

tuning it for non-native tasks. However, these methods still depend heavily on annotated data, perpetuating the limitations of supervised learning.

Our work begins to introduce a novel self-supervised framework called language-adversarial representation learning to overcome these limitations. This framework leverages native speech data from both the learner's first language and the target language for non-native acoustic modeling in mispronunciation verification. First, we design a self-supervised model that learns from target-language speech by predicting future observations within the speech signal. Then, using native-language data, we apply language-adversarial training to align feature distributions between the two languages by training the model to "confuse" a language discriminator.

To enhance verification accuracy and improve feedback quality, we integrate a sinc filter into the self-supervised learning framework. This filter captures formant-like features related to the place and manner of articulation, offering phonetic insights crucial for generating more instructive feedback.

Additionally, we propose methods to enrich the representations learned through self-supervised training for non-native acoustic modeling, including:

- Multi-target contrastive coding: This approach contrasts phonetic discrepancies both within and across languages and speakers, enabling the model to learn nuanced phonetic representations.

- Reconstruction regularization: By recovering the original speech from shared components, this method encourages the model to learn more abstract, transferable features.

Experimental results demonstrate that our approach produces meaningful and transferable representations for non-native acoustic modeling, achieving state-of-the-art performance in non-native phone recognition and mispronunciation verification—all without requiring human supervision.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to give gratitude to everyone who helped me to complete this dissertation.

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Takahiro Shinozaki, for his invaluable guidance, unwavering support, and insightful feedback throughout the course of my research. His mentorship has been instrumental in shaping my academic and personal growth.

I am profoundly grateful to the committee members, for their constructive criticism, encouragement, and thoughtful suggestions, which significantly enhanced the quality of this work.

A special thanks to my family for their unconditional love, patience, and encouragement throughout this endeavor.

Lastly, I would like to acknowledge the countless teachers, mentors, and peers who have inspired and guided me along the way. This work is a testament to the collective contributions of everyone who believed in me.

Thank you all for being part of this journey.

Dedication

This dissertation is dedicated to my family.

Chapter 1

Introduction

With the advancement of globalization, an increasing number of people have come to recognize the importance of mastering a foreign language. Proficiency in a foreign language is often viewed as a critical skill for personal development, academic achievement, and career advancement in today's world. Among the various aspects of language learning, pronunciation stands out as particularly significant. This is because the ultimate goal of learning a foreign language is to communicate with others effectively and efficiently, and spoken language is the most direct and convenient means of achieving this. Natural and accurate pronunciation is essential for successful oral communication, as it ensures that the speaker's message is understood by listeners without confusion or misinterpretation.

In traditional language education, one of the most effective methods for pronunciation training involves face-to-face classroom instruction. In such settings, teachers provide detailed explanations of pronunciation techniques and demonstrate correct pronunciation for learners. Learners then practice by listening to these demonstrations and mimicking the sounds through repeated imitation. During this process, teachers play a crucial role in offering personalized and targeted feedback, addressing the specific pronunciation challenges encountered by each learner. This interactive and guided approach not only helps learners refine their pronunciation skills but also builds their confidence in using the language in real-world scenarios.

This traditional method, while effective, can be resource-intensive and limited in scalability, which underscores the need for innovative approaches to pronunciation training in modern language learning contexts. However, as the demand for language learning grows, traditional teaching methods cannot meet such large-scale needs.

- Firstly, there is a shortage of qualified teachers, and in classrooms with many students, it is difficult for teachers to provide detailed guidance to everyone.
- Secondly, pronunciation practice requires consistent and sustained effort, which cannot be accommodated within the limited time of classroom instruction. Ex-

tending pronunciation training beyond classroom hours is essential. However, when students practice independently without proper methods or teacher feedback, they may engage in repetitive mimicry without making significant progress in pronunciation.

Computer-Assisted Pronunciation Training (CAPT) systems are capable of guiding learners in pronunciation training at both the segmental and suprasegmental levels. At the segmental level, which focuses on vowels and consonants, CAPT systems perform two key tasks: automatic pronunciation assessment and mispronunciation verification. Automatic pronunciation assessment involves evaluating learners' pronunciation based on specific criteria to assign scores or levels. This method provides a direct assessment of overall pronunciation proficiency, making it particularly suitable for examination systems. On the other hand, mispronunciation verification identifies specific phonetic errors in learners' speech and pinpoints the exact types of mistakes, such as mispronouncing a front nasal sound as a back nasal sound. The detailed feedback provided by mispronunciation verification is highly beneficial for learners, enabling them to correct their pronunciation errors in subsequent practice sessions. As a result, this aspect of CAPT has garnered increasing attention in recent years.

However, traditional approaches to mispronunciation verification typically define errors in simplistic categories, such as phoneme-level insertions, deletions, and substitutions. In reality, language learners from different regions often experience varying degrees of native language interference, which results in region-specific pronunciation tendencies. These errors are not merely straightforward phoneme substitutions but rather subtle deviations from standard pronunciation, varying in degree across individuals. Moreover, overly simplistic or negative feedback on errors can significantly discourage learners, reducing their motivation and leaving them uncertain about how to improve their pronunciation.

Traditional acoustic model-based detection systems face significant challenges in addressing these issues. For instance, the performance of such systems is often limited by shortcomings in acoustic modeling and detection algorithms. Further advancements in these areas are necessary to enhance the accuracy and reliability of detection systems. At the suprasegmental level, CAPT systems focus on aspects such as stress and intonation. These features play a critical role in conveying meaning and naturalness in speech, and incorporating them into CAPT systems can provide a more holistic approach to

pronunciation training. As research progresses, the integration of both segmental and suprasegmental feedback is expected to greatly enhance the effectiveness of CAPT systems, empowering learners to achieve more accurate and natural pronunciation.

Mispronunciation verification plays an important role in the CAPT system. Mainstream mispronunciation verification systems are mainly based on the state-of-the-art speech technique that establishes acoustic models for the specified task using relevant speech data (Witt, 1999). It is straightforward to set up the acoustic model of non-native mispronunciation verification using sufficient non-native speech of target language produced by language learners. Many pieces of researches have explored varieties of machine learning techniques for this task. (Duan et al., 2014) explores Gaussian mixture model (GMM) to mispronunciation detection. Benefiting from the deep learning methods, (Gao et al., 2015; Yang et al., 2017) investigate several kinds of deep learning models, including deep neural networks (DNNs), convolutional neural networks (CNNs), and sequence models like recurrent neural networks (RNNs), to mispronunciation detection and achieves encouraging results. However, it is difficult to collect a large amount of non-native speech data. And, annotation is also a time-consuming task since it relies on human speech perception and manual labeling. The data sparsity problem poses large barriers to obtain a high-performance mispronunciation verification system based on supervised learning.

1.1 Overview of Mispronunciation Verification

The task of a mispronunciation verification system is to enable a computer to automatically identify errors in a learner's pronunciation. Currently, the most widely used mispronunciation verification techniques focus on text-dependent mispronunciation verification, where the text to be read by the learner is predefined.

The general process of such systems can be described as follows:

- the system prompts the learner to read aloud a pre-prepared text,
- the learner's speech signal is processed to extract a series of features, such as acoustic or phonetic characteristics.
- the extracted features are then input into the mispronunciation detection module, which compares these features with a pre-trained model, matching them against expected patterns.

- based on the matching results, the system consults a knowledge base containing information about common pronunciation errors.
- the system provides the learner with a detailed report on detected errors, identifying specific mispronunciations and their nature.

This structured process ensures that the system can evaluate the pronunciations of the learners in a systematic and reliable manner. By leveraging acoustic and phonetic features, the system offers precise detection of mispronunciations, enabling targeted feedback. As research in this field advances, the incorporation of more sophisticated feature extraction methods and enriched knowledge bases can further enhance the accuracy and usability of mispronunciation detection systems, paving the way for more effective language learning tools.

Mispronunciation verification techniques can be broadly classified based on the features and modeling approaches they employ into acoustic-phonetic-based detection and speech-recognition-based detection. The core idea of acoustic-phonetic-based detection is to identify distinctive features in speech signals that correspond to specific types of pronunciation errors. This approach relies heavily on acoustic and phonetic knowledge to extract features that are highly discriminative for distinguishing between correct and incorrect pronunciations. Although effective for targeted error types, this method typically requires building specialized classifiers customized to each error category. As a result, computational methods and feature extraction processes often vary between error types, making the integration of such systems into broader CAPT platforms complex and resource intensive.

Given the similarities between mispronunciation verification and speech recognition tasks, many mispronunciation verification systems have adopted techniques from automatic speech recognition (ASR). With recent advances in machine learning, particularly deep neural networks, the performance of ASR systems has improved significantly. Researchers have increasingly applied these advanced ASR techniques to mispronunciation verification tasks. Currently, most mispronunciation verification systems are built within the ASR framework (Abdou et al., 2006). Compared to acoustic-phonetic-based approaches, ASR-based systems offer greater versatility. Instead of requiring separate classifiers for each error type, ASR-based systems can train a single unified model capable of detecting a wide range of pronunciation errors. This reduces complexity and makes the system more scalable. However, the effectiveness of this approach depends

heavily on the precision and robustness of the underlying acoustic modeling, which must be capable of accurately capturing subtle deviations in pronunciation.
This study focuses on ASR-based mispronunciation verification, specifically on the development of acoustic models and learning methods. By leveraging advancements in ASR technology, we aim to improve the accuracy and efficiency of mispronunciation verification systems, making them more effective for real-world applications in CAPT systems. This approach seeks to bridge the gap between the technical demands of mispronunciation verification and the practical needs of language learners.

1.2 Related Works

1.2.1 Early stage

In the early stages of research, confidence-based methods were mainly used due to their relative simplicity in construction and optimization. These methods are language-agnostic, meaning they can be applied to learners of any language regardless of their native language background. The confidence-based approach begins by performing forced alignment between the learner's speech and the corresponding reference text. During the decoding process, intermediate results are generated, and the likelihood probabilities are computed using Hidden Markov Models (HMMs) for evaluation.
One of the earliest studies in this area was conducted by (Kim et al., 1997), who calculated log-likelihood and log-posterior probabilities at the phoneme level using HMMs. The study demonstrated good correlation between speaker-level detection results and human scoring, but phoneme-level evaluation was less reliable.
To address these limitations, (Witt & Young, 2000; Witt, 1999) proposed the Goodness of Pronunciation (GOP) score, a variant of the log-posterior probability. The GOP score evaluates the acoustic features of a learner's speech against phoneme probabilities and uses thresholds to identify errors. This method is straightforward to implement within the ASR framework and does not require knowledge of the learner's native language. (Kanters et al., 2009) further validated the GOP score for Dutch learners, showing its high versatility and minimal sensitivity to speaker variability and threshold selection. Today, the GOP score is widely applied in mispronunciation verification.

1.2.2 Incorporating Linguistic Knowledge

While confidence based methods are effective, they often overlook the influence of native language transfer on second language learning. Many researchers have sought to incorporate linguistic knowledge to address this limitation.

(Y.-B. Wang & Lee, 2012) leveraged inter-language differences to identify error patterns related to a specific native language and developed extended pronunciation dictionaries or recognition networks. These methods demonstrated strong language-specific targeting.

(Liu, 2010) compiled a set of common phoneme-level pronunciation errors based on linguistic analysis, simplifying recognition networks while improving system performance and reducing computational complexity. Using KL divergence between standard and accented models, they enhanced the detection of typical error types.

(lang Wang et al., 2009) combined phonetic knowledge with differences between American English and Cantonese to construct extended pronunciation dictionaries. A pruning algorithm was applied to retain valid pronunciations and eliminate implausible ones, enabling fast and accurate detection of mispronunciations.

1.2.3 ASR-Based Models

In the ASR framework, improving the acoustic model is crucial for enhancing mispronunciation verification. In the era dominated by GMM-HMM models, techniques such as Speaker Adaptive Training (SAT) were employed to address mismatches between acoustic models and speaker characteristics.

For instance, (Ohkawa et al., 2009) used SAT to handle time-frequency differences across speakers, improving the accuracy. (Zhang et al., 2008) extended this by introducing Constrained Maximum Likelihood Linear Regression (CMLLR) for speaker normalization, further mitigating mismatches between training and testing data.

1.2.4 Deep Learning Integration

With the rapid development of deep neural networks (DNNs) (Hinton & Salakhutdinov, 2006; Hinton et al., 2006), ASR systems have achieved significant performance gains, leading to growing interest in deep learning based mispronunciation verification.

(Qian et al., 2010, 2011, 2012) integrated DNNs into CAPT systems, exploring methods

like pre-training and discriminative training to enhance the accuracy.
(Lee & Glass, 2012; Lee et al., 2013) replaced traditional GMMs with DNNs in acoustic models, extracting features from probability matrices to improve detection rates.
(Yuan et al., 2012) used Tandem features derived from DNN outputs, outperforming traditional acoustic models.
(Hu et al., 2013a, 2014a, 2014b, 2015) further applied DNN methods to CAPT systems, significantly improving system performance.

1.2.5 Beyond Simple Error Categories

Traditional definitions of mispronunciation, such as insertion, deletion, or substitution, fail to capture the subtleties of common errors made by second-language learners. Many errors are not straightforward but instead involve slight deviations caused by inaccuracies in articulation placement or manner.
(Cao et al., 2010) introduced the concept of Pronunciation Erroneous Tendencies (PETs), which focuses on deviations in articulation and provides more detailed feedback compared to traditional error categories. PET detection offers valuable insights for second-language learning.
(Duan et al., 2014) explored PET detection using ASR-based methods, demonstrating its feasibility.
(Gao et al., 2015) replaced GMM-HMM models with DNN-HMM hybrids, improving PET detection accuracy.

1.2.6 Addressing Data Sparsity

Annotating PETs requires expertise in phonetics and is labor-intensive, making it challenging to obtain large annotated datasets. Insufficient training data often leads to overfitting and poor model performance.
(Duan et al., 2019) tackled this issue by using multilingual training data (e.g., Japanese and Chinese) and multi-task learning, achieving improved the performance in PET detection tasks for Chinese learners of Japanese.

1.2.7 Summary

Advances in ASR based frameworks, particularly with the integration of deep learning

and linguistic knowledge, have significantly improved mispronunciation verification systems. However, challenges such as data sparsity and nuanced pronunciation deviations remain critical areas for further research and innovation.

1.3 Mispronunciation Definition

The traditional definition of pronunciation errors typically categorizes them into simple types such as insertion, deletion, or substitution. However, extensive research has shown that common pronunciation errors encountered by foreign second-language learners when learning Mandarin cannot be easily classified into these simple categories. Their pronunciation errors often involve deviations from standard pronunciation, caused by inaccuracies in their articulation methods and positions. Additionally, pronunciation errors vary among second-language learners from different countries, influenced by the phenomenon of native language transfer. As a result, their pronunciation often exhibits distinct regional characteristics, and these errors cannot simply be defined as substitution errors. According to the definition proposed by (Cao & Zhang, 2009), focusing on slight inaccuracies in articulation placement and manner. Compared to traditional definitions of pronunciation errors such as insertion, deletion, or substitution, pronunciation deviation trends provide more precise and detailed information about mispronunciations. We represent these types of pronunciation deviations using a large number of phonetic symbols for annotation. These deviation types include nasalization shifts, tongue height variations, aspiration length differences, lip rounding or spreading, and more. All deviation types are defined based on inaccuracies in articulation position and manner. Parts of definitions can be found in Table 1. The total phone set is set up to combine the standard pronunciations and the defined mispronunciations, total size is 267 including 65 mispronunciations.

1.4 Dissertation Research and Contributions

In this dissertation, we propose a series of self-supervised framework aiming to address the data sparsity problem, which is the main problem, to improve the performance of mispronunciation verification system. we have proposed several novel framework for mispronunciation and the contributions can be summarized as follows:

Table 1: Parts of mispronunciation definitions.

Errors	Diacritics	E.g.	Notation
Spreading	w	u{w}	The round sound /u/ has a spreading lip.
Rounding	o	e{o}	The spreading lip sound /e/ is pronounced to the round sound.
Backing	-	n{-}	The tongue position of the phoneme /n/ is a little back.
Advancing	+	e{+}n	The tongue position of the phoneme /e/ is a little advancing so the pronunciation of /en/ is like that of /n/.
Shortening	;	p{;}	The aspiration duration of the phoneme /p/ is a little shorter.
Lengthening	:	z{:}	The fricativizing duration of the phoneme /z/ is a little longer.
Laminalizing	sh	sh{sh}	The balade-palatal phoneme /sh/ is pronounced like Japanese limina-alveolar

- We propose a self-supervised framework for non-native mispronunciation verification. We expect that the knowledge learned via this framework from a large scale of speech data from two native languages can help relieve the data sparsity problem for non-native mispronunciation verification. In our work, a large scale of unlabeled raw speech from the target language is fed to the model to capture phonetic properties by make predictions about the observation in the speech of the target language. Then the model is trained with language adversarial training using the learner's native language. In this manner, we expect the model can capture some kinds of patterns between two languages.

- Based on the proposed framework above, we introduce sinc filter to extract formant-like features. Formant is considered relevant to some kinds of mispronunciations from the respect of placements and manners of articulation Wu and Lin, 1989. Since non-native pronunciation is easily affected by the learners' native language, which is known as L1 transfer Iverson et al., 2003, and some kinds of pronunciations errors are sort of deviations from the canonical ones. This information is useful not only for detecting pronunciation errors but also to be able to provide detailed and instructive feedback to guide the learners to correct their pronunciation errors.

- We explore to enrich the representations learned by self-supervised training for

non-native acoustic modeling, mainly including two schemes: 1) we propose multi-target contrastive coding. Several researches have indicated that the predictive coding-based approaches contrastive to different targets can capture different information, e.g., the model with negative samples from the same utterance to the input can capture phonetic structure Oord et al., 2018, and which with the targets from different speakers' utterances can learn speaker-related information Mirco and Yoshua, 2019. Inspired by it, in this work, our model is designed to make predictions about the observations contrastive to different targets jointly to learn the representations of the discrepancy with respect to phonetic structures in and across languages, and speakers at the same time. 2) an additional term to reconstruct the original speech from the shared components. This term serves as a regularization that leads the intermediate representations learned by the model to be a good abstraction of the input speech. With these two schemes, we expect that the representations learned by our propose approaches from two native languages will be transferable and meaningful for non-native acoustic modeling of phone recognition and mispronunciation verification without human supervision.

Chapter 2

Language Representation Learning for Mispronunciation Verification

This chapter presents the language representation learning for mispronunciation verification. The work presented in this chapter has been published in INTERSPEECH 2020 (Yang et al., 2020) and Neural Networks Journal (Yang, Fu, Zhang, & Shinozaki, 2021).

2.1 Introduction

The growing demand for second language (L2) acquisition, driven by globalization and increasing social integration, has heightened interest in computer-aided language learning (CALL) systems in recent years. Compared to traditional one-on-one communicative approaches between teachers and students, CALL systems offer greater flexibility and significantly reduce resource requirements, including teaching staff, administrative personnel, and physical classroom spaces.

As a key component of Computer-Assisted Language Learning (CALL) systems, Computer-Aided Pronunciation Training (CAPT) serves as a virtual teacher that processes and analyzes learners' utterances, providing assessments of their pronunciation quality (Hu et al., 2013b; Witt & Young, 2000; Zheng et al., 2007). CAPT systems are expected not only to accurately detect pronunciation errors in non-native speech (Harrison et al., 2009; Y.-B. Wang & Lee, 2012) but also to diagnose the types and locations of these errors. Furthermore, CAPT should offer constructive feedback to help learners improve their pronunciation in future attempts (Hu et al., 2015; Wei et al., 2009). This capability is referred to as mispronunciation verification. For instance, if a student pronounces the syllable "u" as "uw"—indicating that their lips spread instead of forming the rounded shape required for the sound /u/—the system should do more than simply flag the error with a message like "You have a pronunciation error." It should also provide specific guidance, such as explaining how to produce the correct sound. Research has shown that incorporating information about the place and manner of articulation in feedback

can effectively help learners form a clear and detailed understanding of how to adjust their articulation (Jo et al., 1998; Koreman et al., 2013). For the example above, the system could suggest: "Try not to spread your lips when producing the rounded sound /u/." This type of precise, actionable feedback mirrors the guidance a professional language teacher might provide, making the learning process more efficient and impactful.

Mispronunciation verification is a critical component of CAPT systems. Mainstream mispronunciation verification methods primarily leverage state-of-the-art speech processing techniques, which involve building acoustic models tailored to the specific task using relevant speech data (Witt, 1999). A straightforward approach is to develop an acoustic model for non-native mispronunciation verification by utilizing a sufficient amount of non-native speech data in the target language, produced by language learners. Numerous studies have investigated various machine learning techniques for this purpose. For example, (Duan et al., 2014) explored the application of Gaussian Mixture Models (GMMs) for mispronunciation detection. Leveraging the advancements in deep learning, several studies (Gao et al., 2015; Yang et al., 2017) have explored various deep learning models for mispronunciation verification, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and sequence-based models such as Recurrent Neural Networks (RNNs). These approaches have yielded promising results, demonstrating the potential of deep learning in this domain.

However, collecting a large volume of non-native speech data presents significant challenges. Additionally, the annotation process is time-consuming, as it depends on human speech perception and manual labeling. This issue of data sparsity creates substantial obstacles to developing a high-performance mispronunciation verification system using supervised learning.

To address the data sparsity problem in non-native acoustic modeling, many approaches have been investigated. These investigations can be roughly grouped into two directions.

- One is to explore some varieties of techniques trying to efficiently utilize limited non-native speech data.

  (Gao et al., 2016) proposed a method to diversify input for non-native acoustic modeling. They first developed an articulatory features (AF) model targeting the places and manners of articulation. The bottleneck features, extracted from

the intermediate layers of the AF model, were then combined with the original acoustic features to train a non-native mispronunciation detector.

(Yang et al., 2017) investigated data augmentation techniques to mitigate the data sparsity issue. Their approach involved corrupting clean training speech with noise and applying various perturbations to increase the quantity and diversity of training data. Results demonstrated that combining clean data with perturbed data effectively improved performance by enlarging the scale of non-native training data.

lin2020improving explored integrating multiple models into a unified system using soft targets. This approach involved employing various models to generate hidden representations and then ensembling them to make final decisions, thereby leveraging the strengths of different models to enhance performance.

- Another direction is based on transfer learning. Most of these approaches firstly establish a model using a large amount of data on a general task, such as speech recognition, to capture features of relevant properties and then apply these features to non-native tasks.

  (Bouselmi et al., 2006; Joshi et al., 2015; Lee & Glass, 2015; Uebler & Boros, 1999; Z. Wang et al., 2003) explore many kinds of approaches based on the multilingual framework for non-native acoustic modeling for mispronunciation detection.

  (Duan et al., 2019) explore multi-lingual framework using two native corpora, including language learner's native language and target language, to transfer the phonetic and articulatory knowledge learned with multi-task architecture to non-native mispronunciation detection.

However, most of these approaches still rely on supervised learning that demands plenty of training data as well, and, specifically, corresponding annotations that are time-consuming to be obtained either.

In recent years, several studies have introduced promising self-supervised approaches to obtain speech representations (Hyvarinen & Morioka, 2016; Oord et al., 2018; Rivière et al., 2020). For instance, (Schneider et al., 2019) proposed the contrastive predictive

coding (CPC) model, which allows for learning various speech properties, such as phonetic content and speaker characteristics, from a large amount of unlabeled data without the need for annotations. These learned properties can then be applied to downstream tasks, such as automatic speech recognition and speaker verification (Kawakami et al., 2020). Compared to supervised learning approaches, this method is more scalable and cost-effective, as it does not require labeled data.

Inspired by these advantages, we argue that it would be more efficient and flexible to leverage knowledge learned through self-supervised approaches from large amounts of raw speech data in both the learner's native and target languages. Since these data are relatively easy to collect, this approach could help address the data sparsity issue in mispronunciation verification.

In this chapter, we propose a self-supervised learning-based pre-training approach for non-native mispronunciation verification. We hypothesize that knowledge acquired from large-scale, unlabeled speech data in two languages—native and target—can help alleviate the data sparsity problem for non-native mispronunciation verification. Specifically, a large corpus of unlabeled raw speech from the target language is used to capture phonetic properties by making predictions about the observed speech. The model is then trained with language-adversarial techniques using the learner's native language. This approach aims to help the model learn patterns shared between the two languages. Since non-native pronunciation is often influenced by the learner's native language (a phenomenon known as L1 transfer) (Iverson et al., 2003), we expect the model to capture pronunciation errors as deviations from canonical forms.

This work investigates two kinds of approaches to perform adversarial training:

- an explicit auxiliary task, in which a language discriminator is introduced to distinguish which language the input sample is
- implicitly scheme in which sampling from learner's native data as negative samples when calculating the loss of prediction for the target language.

Then the pre-trained model is applied as a part of the downstream mispronunciation verification task. We evaluate the performance of our proposed approaches to the mispronunciation verification task on the Japanese part of the BLCU inter-Chinese speech corpus, which is designed and collected for the learners from Japan, who learn Man-

darin Chinese as their second language.

2.2 Language Adversarial Representation Learning

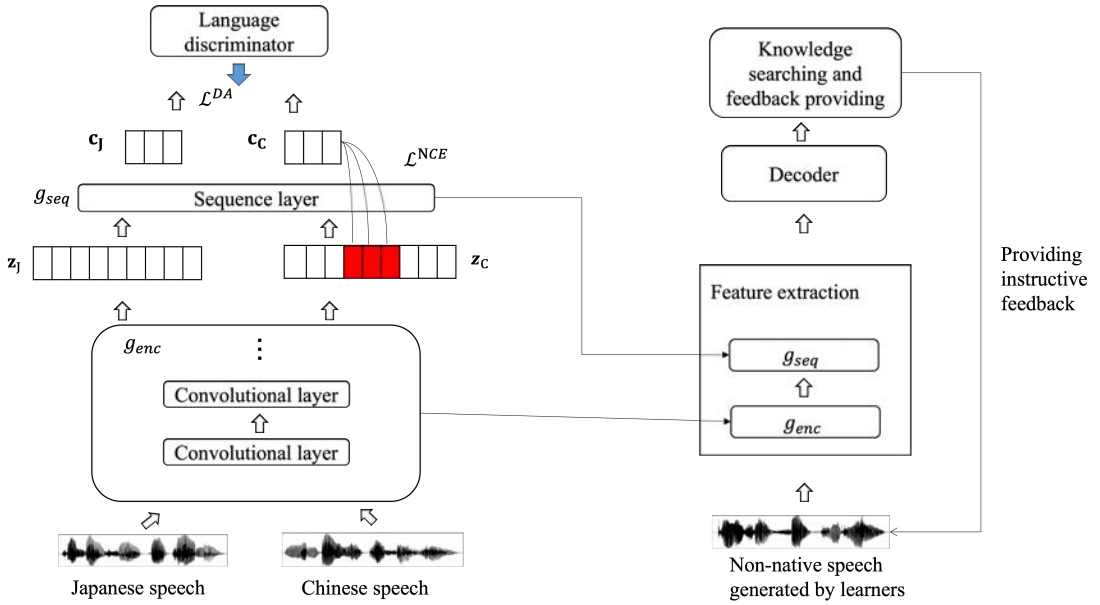

Figure 1: A demonstration of our proposed language adversarial representation learning framework.

Our goal is to extract representations of knowledge from raw speech data in two native languages for non-native acoustic modeling without human supervision. In our framework, the speech data of the target language (Chinese) is input into a self-supervised pre-trained model to capture phonetic structure information. The model is then trained using language-adversarial techniques, where speech from the native language (Japanese) is fed into the model to confuse a language discriminator, which distinguishes the language of the input. This adversarial training aligns the feature distributions of the two languages. We first explain the pre-training process with the pre-trained model, followed by an introduction to language-adversarial training. Finally, we describe the mispronunciation verification framework that incorporates the pre-trained model.

Figure 1 illustrates our proposed language-adversarial representation learning framework. In the pre-training stage, shown on the left, the model receives target language (Chinese) data to capture phonetic structure in an unsupervised manner. Simultaneously, speech from the learner's native language (Japanese) is input to align the feature distributions of the two languages. The figure shows language-adversarial training

with an explicit auxiliary task, where the model is paired with an additional output that shares hidden representations from the target language input. It is important to note that Japanese speech data do not contribute to calculating the model's loss $\mathcal{L}^{NCE}$, but only participate in the adversarial loss $\mathcal{L}^{DA}$ for the language discriminator. The solid arrow indicates that the loss is passed through a gradient reversal layer to confuse the language discriminator. After pre-training, the pre-trained model is integrated into the downstream mispronunciation verification framework.

2.2.1 Encoder

Our pre-trained model is based on an unsupervised representation learning framework utilizing slow feature analysis (Oord et al., 2018). The objective is to extract features that make long-term predictions about future observations while preserving the properties and structures of the input. This is achieved by maximizing the mutual information between the features and those extracted from future timesteps. Predictions across different timescales capture varying levels of information: rapidly changing representations are indicative of local structures, while more slowly varying ones correspond to higher-level abstractions or global structures, such as phonemes and words in the speech signal (Wiskott & Sejnowski, 2002).

An overview of the model is illustrated in the solid block in the left part of Figure 1. In detail, let $\mathbf{x} = \{x_1, x_2, ..., x_l\}, x_i \in \mathbb{R}$, denotes an raw speech signal in L discrete time steps where x_i is the acoustic amplitude at time i. First, an encoder g_{enc} encodes the signal into embedding vector representations $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_t\}, \mathbf{z}_i \in \mathbb{R}^{d_z}$, where d_z is the dimension of the hidden representation.

In the processing of encoding, the inputs are first fed into convolutional layers to generate embedding vector representations $\mathbf{z}$, which denotes the hidden representations of speech signal.

The brief process of the encoder g_{enc} can be written:

$$\mathbf{z} = g_{enc}(x_1, x_2, ..., x_l) \tag{1}$$

Then a sequence model g_{seq} summarizes the past information of vector and produces corresponding context-aware representations, which can be denoted as $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t\}, \mathbf{c}_i \in$

$\mathbb{R}^{d_c}$, i.e.,

$$\mathbf{c} = g_{seq}(\mathbf{z}), \tag{2}$$

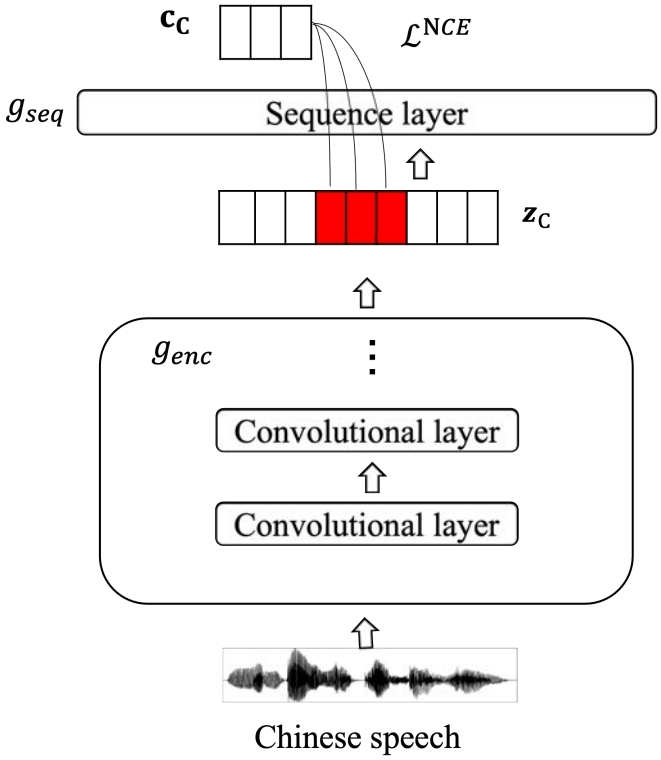

Figure 2: A demonstration of calculation of InfoNCE.

The optimization is performed by minimizing the InfoNCE (Oord et al., 2018), which is a loss function based on noise contrastive estimation as the lower-bound to the mutual information between context aware embedding $\mathbf{c}_t$ and future latent representations $\mathbf{z}_{t+k}$ for $k \in \{1, ..., K\}$. Given a set $Z = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ which contains one positive sample from $p(\mathbf{z}_{t+k}|\mathbf{c}_t)$ and $N-1$ negative samples from "noise" distribution $p(\mathbf{z})$. The calculation of InfoNCE can be described in Figure 2, which demonstrates the process of calculation of InfoNCE. The InfoNCE loss function for each step t can be donated as follows:

$$\mathcal{L}_{tk}^{NCE} = -\mathbb{E}\left[\log \frac{f_k(\mathbf{c}_t, \mathbf{z}_{t+k})}{\frac{1}{N}\sum_{\tilde{\mathbf{z}} \in Z} f_k(\mathbf{c}_t, \tilde{\mathbf{z}})}\right] \tag{3}$$

where $f_k(\mathbf{c}_t, \mathbf{z}_{t+k})$ is a scoring function that can be a lob-bilinear model:

$$f_k(\mathbf{c}_t, \mathbf{z}_{t+k}) = \exp(\mathbf{c}_t^T \mathbf{W}_k \mathbf{z}_{t+k})) \tag{4}$$

where $\mathbf{W}_k$ is the parameters in each model for each k.
The total loss to be minimized is a sum of the InfoNCE loss for each step:

$$\mathcal{L}^{NCE} = \sum_t \sum_k \mathcal{L}_{tk}^{NCE} \tag{5}$$

in which negative samples are sampled uniformly from representations in the same speech signal $\mathbf{z}$.

2.2.2 Language adversarial training

Intuitively, identifying as many patterns as possible in non-native speech is crucial for developing an effective mispronunciation verification model. However, the data sparsity problem complicates this task, as non-native pronunciation datasets are typically small, and the proportion of pronunciation errors is generally much lower than that of correct pronunciations.
A key challenge lies in capturing similar patterns when using speech data from both the learner's native and target languages. We hypothesize that non-native patterns in the target language speech can be influenced by the native language. By using language-adversarial training to align the hidden representations, we aim to capture these non-native patterns. This process mirrors language learning, where a learner's target language speech is often influenced by their native language, resulting in pronunciation errors that are typically small deviations from canonical forms. The goal of incorporating the language-adversarial component is to capture such non-native patterns by leveraging speech data from both languages.
Two schemes of language adversarial training are investigated.

- One is to explicitly introduce and confuse a language discriminator to perform language discrimination as an auxiliary task.
- Another is to sample negative samples from speech data of different languages when calculating the InfoNCE loss.

In below, we describe these two approaches.

2.2.2.1 Explicit language adversarial training with auxiliary task

The proposed language adversarial with the auxiliary task is shown in Figure 1. In this setup, the pre-trained model is paired with another output that shares the internal representations of the input and tries to discriminate whether the input speech signal comes from the target language or not. In this experiment, we focus on Chinese speech produced by language learners from Japan (the native is Japanese $\mathbf{x}_J$ and the target is Chinese $\mathbf{x}_C$). But the proposed approach is not limited to Chinese-Japanese conditions. The training process of this language discriminator is adversarial with respect to the shared hidden layers by using gradient reversal to maximize the language adversarial loss rather than minimize it to confuse the discriminator. The language discriminator is optimized using the negative log-probability as the language adversarial loss:

$$\mathcal{L}^{LA} = -l \log y_w - (1-l)\log(1-y_w) \tag{6}$$

$$y_w = p(l=1|\mathbf{h};\theta_{LA}) = softmax(W_l\mathbf{h}+b) \tag{7}$$

where $l \in \{0,1\}$ denotes the ground-truth language label of input, y_w and $W_l, b \in \theta_{LA}$ are the output and weights of the final layer, and **h** is the hidden representation of the model [1].

In our experiments, we make a comparison between using the output of convolutional layers and that of sequence layers as the input of the language discriminator.

We optimize the weighted-sum of two loss functions using a hyper-parameters λ. The overall training objective of the composite model can be written as follows:

$$\mathcal{L} = \lambda \sum_N \mathcal{L}^{NCE}(\mathbf{x}_C) - (1-\lambda)\left[\sum_N \mathcal{L}^{LA}(\mathbf{x}_C) + \sum_M \mathcal{L}^{LA}(\mathbf{x}_J)\right] \tag{8}$$

where N is the number of target language (Chinese) $\mathbf{x}_C$ and M is that of learner's native language $\mathbf{x}_J$ ($M = N$ in this work). Note that in this manner, we sample negative point from the same utterance with the input.

[1] Note here **h** is used as a collective name, **h** can be the output of encoder **z** or the context vector **c**. We make the comparative study later.

We look for parameters that satisfy a min-max optimization criterion as follows:

$$\min_{\theta_{enc}} \max_{\theta_{LA},\theta_s} \mathcal{L} \tag{9}$$

where θ_{enc}, θ_{LA} and θ_s denote parameters of the encoder, language discriminator and shared part respectively. Such optimization will involve a maximization with respect to the language discriminator and a minimization with respect to the refined pre-trained model.

Algorithm 1: Language adversarial training

Input: Chinese data $\mathbf{x}_C$, Japanese data $\mathbf{x}_J$, batch size b

Output: learned model parameters

1. Initialize model parameters;

2. **repeat**

(1) Randomly sample $\frac{b}{2}$ examples from $\mathbf{x}_C$

(2) Randomly sample $\frac{b}{2}$ examples from $\mathbf{x}_J$

(3) Compute $\mathcal{L}^{NCE}$ and $\mathcal{L}^{LA}$

(4) Take a gradient step for $\lambda\frac{2}{b}\nabla_{\theta_{enc}}\mathcal{L}^{NCE}(\mathbf{x}_C)$

(5) Take a gradient step for $(1-\lambda)\frac{2}{b}\nabla_{\theta_{LA}}\mathcal{L}^{LA}(\mathbf{x}_C)$

`// Gradient reversal`

(6) Take a gradient step for $-(1-\lambda)\frac{2}{b}\nabla_{\theta_s}\mathcal{L}^{LA}(\mathbf{x}_J)$

until convergence;

Algorithm 1 presents pseudocode for the language adversarial training to train the model. The parameters are initialized first. Then we create mini-batches by randomly sampling $b/2$ samples from $\mathbf{x}_C$ and $b/2$ from $\mathbf{x}_J$. Note that, Chinese samples take part in calculating both $\mathcal{L}^{NCE}$ and $\mathcal{L}^{LA}$ while Japanese data only participate in computing the $\mathcal{L}^{LA}$.

2.2.2.1 Implicit language adversarial training with auxiliary task

We investigate another way to incorporate language adversarial training. In the previous section, The $N-1$ negative samples are selected from the same utterance with the input to calculate the InfoNCE loss as mentioned in Section 2.2.1. The main dif-

ference from that is, in this scheme, those $N - 1$ negative points are selected from the Japanese utterance to obtain the loss rather than selecting negative samples from the same utterance of the input. The advantage of this approach is that there are not many additional parameters and loss to be introduced. The comparative results and discussion are discussed later.

2.3 Experiments

2.3.1 Datasets

Native corpora

AISHELL corpus is employed as the Mandarin Chinese source, which is an open-source Mandarin Chinese speech corpus (Bu et al., 2017). And Corpus of Spontaneous Japanese (CSJ) is used as the Japanese source, which is a database containing a large collection of Japanese spoken language data (Maekawa et al., 2004). We randomly choose 300 hours of data from two corpora (150h from Chinese, 150h from Japanese) above as our training set for pre-training.

Non-native corpus

BLCU inter-Chinese speech corpus, which is collected for language learners who learn Mandarin Chinese as their second language (Cao et al., 2010), is employed as our non-native dataset for mispronunciation verification. This work focuses on the Japanese part. Table 2 present the detail about the used dataset. It contains speech from 17 Japanese speakers. Each speaker generate 301 utterances in mandarin and totally 4,631 utterances involving 64,190 phonemes are included. Recordings are made in the sound-proofing speech lab with the sampling rate of 16kHz then encoded in 16-bit pulse-code modulation (PCM). All ground-truth labels in this corpus are annotated according to the discussion of well-trained phoneticians. Around 80% of this corpus is used as the training set, 10% is used as the developing set and the rest for testing. There is no overlap of speakers between the training and testing set and leave-one-out cross-validation (Browne, 2000) is adopted.

Table 2: The detail of non-native dataset.

Text	301
Speakers	7
Number of utterances	1899
Number of phonemes	26431
Average length per utterance	14
Number of annotators	6
Number of annotators per utterance	2

2.3.2 Experimental setup

Non-native only

We first establish the mispronunciation verification framework using only non-native speech data as one of our baselines. We establish several models explored in previous researches (Gao et al., 2015; Lin et al., 2020; Yang et al., 2017) including

- deep feed-forward neural network (DNN) with five layers where each layer has 550 units,
- convolutional neural network (CNN) with three convolutional layers where the size of layers are [128, 256, 512] and size of filters are [3, 3, 3] sequentially,
- bi-directional recurrent neural network with gated recurrent units (GRU-RNN) with one layer of 550 units (Cho et al., 2014) for acoustic modeling.

The 40-dimensional surface feature Mel-frequency cepstral coefficients (MFCCs) extracted from non-native data through 25ms windows with a 10ms frameshift is employed as the input feature. All the feature are applied cepstral mean and variance normalization (CMVN) (Viikki & Laurila, 1998) before training. Batch normalization (Ioffe & Szegedy, 2015) and a dropout (Srivastava et al., 2014) of 0.2 are employed following each layer. RmsProp (Tieleman & Hinton, 2012) is employed as the optimizer with a batch size of 16.

Pre-trained part

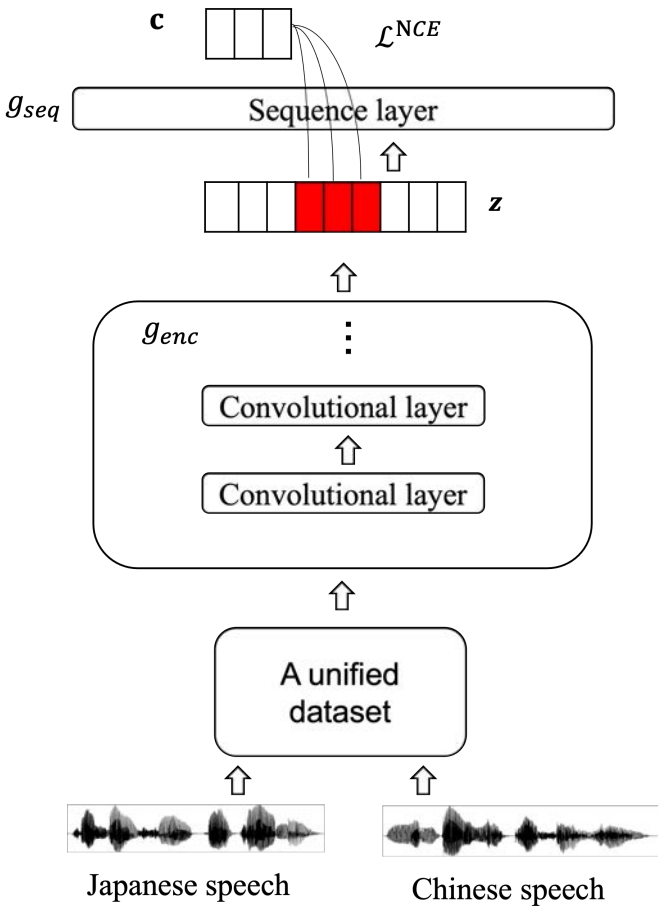

Figure 3: A demonstration of pre-trained model w/o LAT.

Pre-trained model w/o LAT. One of the baselines for pre-training approaches is the pre-trained model without LAT using multi-lingual data. Figure 3 demonstrates the pre-trained model without LAT. We mix the data from two languages (Mandarin Chinese and Japanese) into a large multi-lingual dataset for training the pre-trained model. The encoder contains five 1-dimensional convolutional layers with a 160 down-sampling factor thus there is a feature vector for every 10ms of speech, which keeps consistent with the rate of phoneme sequence labels obtained with Kaldi. For convolutional layers, the size of filters are [10, 8, 4, 4, 4], the strides are [5, 4, 2, 2, 2] and the paddings are [3, 2, 1, 1, 1]. 512 hidden units of each layer are with ReLU activation. Batch normalization is employed following each convolutional layer. A recurrent neural network with gated recurrent units (GRU-RNN) with 256-dimensional hidden state is employed as the sequence model. The output of GRU at every timestep is used as the context $\mathbf{c}$ to predict 12 timesteps in the future. In each training iteration, a segment containing 20480 data points (around 1.28s) is randomly selected from the speech for every utter-

ance. Adam optimizer (Kingma & Ba, 2014) with a learning rate of 2e-4 is used to train the model with a minibatch whose size is 8.

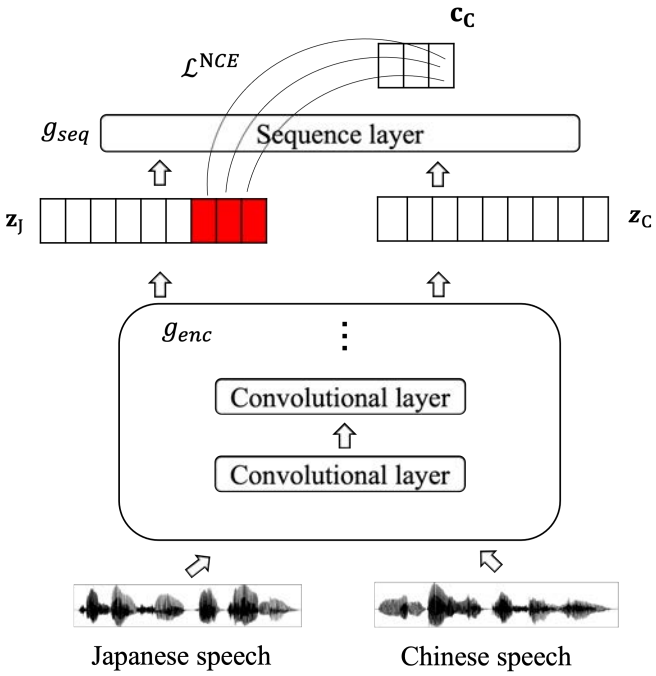

Figure 4: A demonstration of implicit LAT.

Implicit LAT. For implicit language adversarial training, there is no additional language discriminator. Figure 4 demonstrates the implicit LAT setting. The main difference from the pre-trained model w/o LAT and following explicit language adversarial training is that, when calculating InfoNCE, the negative samples are selected from the learner's native language (Japanese in this work), which reflects the idea of language adversarial.

We explore two explicit language adversarial training approaches shown as follows:

Shallow explicit LAT. In this approach, we take the output of the last convolutional layer as the input into the language discriminator. Figure 5 present the shallow explicit LAT. The language discriminator contains two layers with hidden units whose size is [64, 2], and the final layer output the one-hot representations of two languages. λ in Eq (9) are set to 0.5.

Deep explicit LAT. The main difference from deep explicit lat is that the output of the sequence model is fed into the language discriminator. Figure 6 depict the deep explicit LAT. The configuration of language discriminator is similar to that of shallow explicit LAT.

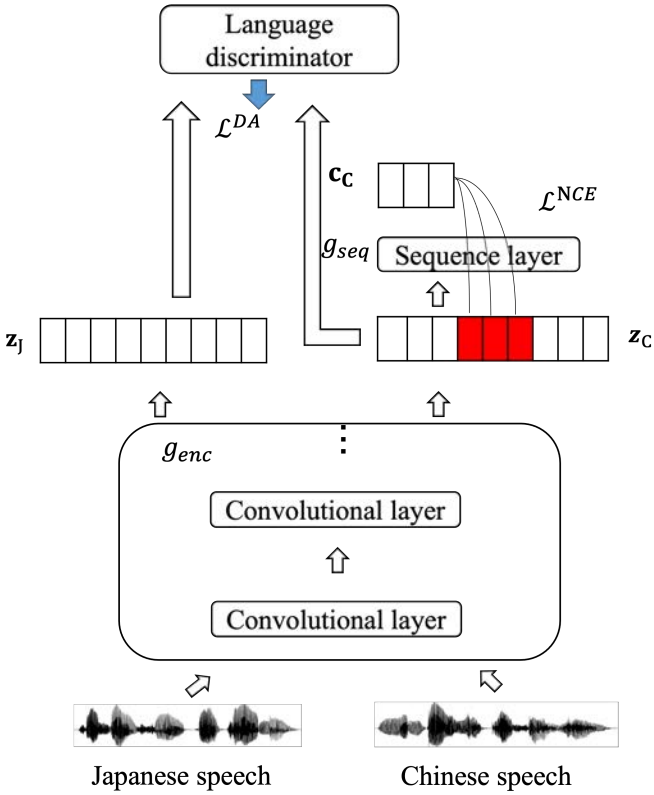

Figure 5: A demonstration of shallow explicit LAT.

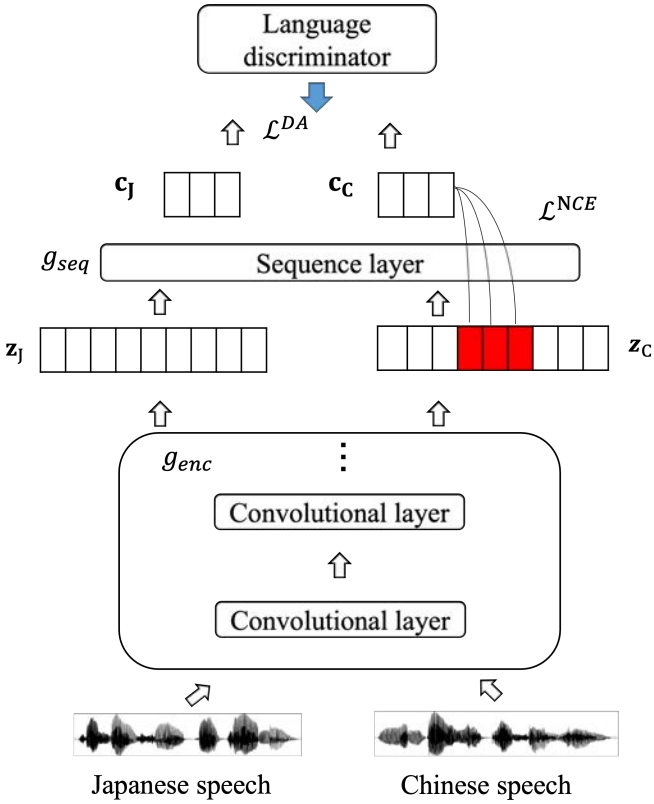

Figure 6: A demonstration of deep explicit LAT.

Downstream non-native part

Then we employ our pre-trained model in the mispronunciation verification framework in which the pre-trained model serve as a feature extractor. The output, 256-dimensional vector representation, of the pre-trained model is directly sent to the downstream decoder (Note that with pre-trained models, only the vector representations are employed as features). The mispronunciation verification framework is a hybrid neural network system. The senone labels (tied HMM states) are first obtained by a Gaussian mixture model-hidden markov model (GMM-HMM), then these labels and the corresponding aligned frames are used for training. The same align information is employed by all schemes. The decoder use one layer of bi-GRU with 550 units. Batch normalization and a dropout of 0.2 are performed following each layer. RmsProp is employed as the optimizer with a batch size of 16. The configurations are the same for all of the pre-training approaches for a fair comparison. Whether the parameters of the pre-trained model is trainable is explored in the experiments.

2.3.3 Evaluation metrics

2.3.3.1 Metrics for non-native phone recognition

The CAPT system must first be capable of recognizing non-native speech in high performance. In this work, our primary focus is on evaluating the performance at the phone level, encompassing both standard pronunciations and predefined mispronunciations. A key motivation for employing phone error rate (PER) as a metric is its ability to assess the system's performance in recognizing all phones. It is important to note that mispronunciations constitute only a small fraction of all phones and are inherently a subset of the total phone set. This imbalance underscores the criticality of accurate recognition, as any errors in identifying correct pronunciations as mispronunciations could lead the model to generate misleading feedback for learners. Such inaccuracies would compromise the system's reliability and tamp down learners' enthusiasm. Therefore, achieving robust overall performance across all phonemes is essential to ensure the CAPT system provides consistent and constructive feedback.

The PER can be defined as follow:

$$PER = \frac{S + D + I}{N} \tag{10}$$

where N is the total number of all the phones. S, D and I are the numbers of substitution, deletion and insertion respectively.

2.3.3.2 Metrics for non-native mispronunciation verification

The recall and precision are employed as the evaluation metrics for mispronunciation verification. The $recall$ measures that, among all of the phones labeled as the errors manually, how many errors are detected by the detection system. The $precision$ measures that how many mispronunciations detected by the system are truly pronunciation errors. $F1 - score$ is used since we consider that the precision and recall are equally important for the language learners when using CAPT systems. DA (detection accuracy) measures the overall performance of mispronunciation verification. Among all of the 65 kinds of mispronunciations that occurred in the corpus, 16 most common errors were selected for analysis.

$$Recall = \frac{TR}{TR + FA} \tag{11}$$

$$Precision = \frac{TR}{TR + FR} \tag{12}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{13}$$

$$DA = \frac{TA + TR}{TA + TR + FA + FR} \tag{14}$$

Where TR (true rejection) notes the number of phones labeled as the errors by the expert at the same time detected as the errors by the detection system. TA (true acceptance) denotes the number of phone segments that are marked as the correct pronunciation by the system and the ground truth. FR (false rejection) refers to the number of phones recognized as pronunciation errors by the system while the ground truths are correct. FA (false acceptance) are the number of phone segments that are misrecognized as correct while they are actually errors.

2.4 Results

2.4.1 Results of Non-native phone recognition

To evaluate the performance, the system should recognize phones besides that not defined as mispronunciations first. Table 3 demonstrates the detection performance of phone recognition with different approaches. “Non-native only” denotes that the model we directly train on only non-native data. “Implicit LAT” is the implicit language adversarial training with negative sampling from target language. “Shallow explicit LAT” denotes we use the output of encoders as the input fed into language discriminator and ”Deep explicit LAT” means we use the output of sequence model for language discriminator. “fine-tune” denotes that the parameters of the pre-trained model are trainable, in contrast to that these parameters are frozen in other case, when downstream task is performed. The best results are marked with bold fonts. The result marked as bold is shown a statistically significant improvement to baselines at the 0.01 level using Wilcoxon signed-rank test.

Non-native data only vs. Pre-trained model w/o LAT

As shown in Table 3, for overall performance, it can be found that the performances of using the pre-trained model are better than that of using non-native data only, which demonstrates that the features learned from two native speech via an unsupervised approach are useful for non-native spoken language processing. Among them, the pre-trained model w/o LAT using multilingual data achieves slight improvement.

Pre-trained model w/o LAT vs. Pre-trained model w/ LAT

By introducing language adversarial training, our proposed Implicit LAT and Deep explicit LAT outperform the pre-trained model w/o LAT with the phone error rate (PER) from 10.78% to 10.37%, and 10.78% to 9.85%. In the process of training the pre-trained model w/o LAT with multilingual data, the negative sampling is performed in the same utterance with the input for each language separately. It may lead that too many Japanese native patterns are involved, which is not what we expect since our goal is to set up acoustic modeling for non-native speech data that is the utterance of the

target language generated by the non-native learner.

Comparison among different LAT settings

It can be also noticed that, among explicit and implicit schemes of language adversarial training, the performance by using the output of the sequence model Deep explicit LAT as the input of language discriminator, PER of 9.85%, is better than Shallow explicit LAT with PER of 11.03% using that of the encoder. We think that it may be because the sequence layer in Shallow explicit LAT does not receive the guidance from language information through back-propagation from the language discriminator since the output of the encoder is sent as the input of the language discriminator.
Meanwhile, incorporating language adversarial learning with negative sampling from the different language Implicit LAT is better than Shallow explicit LAT but not exceed Deep explicit LAT with an explicit auxiliary task. We think it is because the in-language information may be missing to a certain degree by a single simple task in which one positive of the target language and amount of native language are utilized contrastively.

Frozen parameters vs. Fine-tuning

It also can be found that by fine-tuning the parameters in the Deep explicit LAT, the performance can be further improved from 9.85% to 9.73%. Finally, the best performance with PER of 9.73% is obtained by Deep explicit LAT with fine tune and it outperforms the models directly trained on non-native data only.

Table 3: Detection performance of phone recognition for different approaches.

Model	PER
Non-native only	
DNN	12.25%
CNN	12.09%
GRU	11.22%
Proposed pre-trained approaches	
Pre-trained model w/o LAT	10.78%
Implicit LAT	10.37%
Shallow explicit LAT	11.03%
Deep explicit LAT	9.85%
Deep explicit LAT fine-tune	**9.73%**

Table 4: Detection performance of mispronunciation verification for different approaches.

Model	Recall	Precision	F1 score	DA
Non-native only				
DNN	38.97%	48.14%	43.07	82.04%
CNN	40.35%	48.88%	44.2	82.55%
GRU	40.12%	54.92%	46.37	84.66%
Proposed pre-trained approaches				
Pre-trained model w/o LAT	34.73%	54.2%	42.33	84.29%
Implicit LAT	41.92%	50.72%	45.90	84.88%
Shallow explicit LAT	37.72%	51.79%	43.64	84.79%
Deep explicit LAT	44.91%	58.14%	50.68	86.44%
Deep explicit LAT w/ fine-tune	**45.3%**	**58.5%**	**51.06**	**86.61%**

2.4.2 Results of non-native mispronunciation verification

Table 4 demonstrates the detection performance of non-native mispronunciation verification with different approaches. ”Non-native only” denotes that the model we di-

rectly train on only non-native data. “Implicit LAT” is the implicit language adversarial training with negative sampling from target language. “Shallow explicit LAT” denotes we use the output of encoders as the input fed into language discriminator and “Deep explicit LAT” means we use the output of sequence model for language discriminator. “fine-tune” denotes that the parameters of the pre-trained model are trainable, in contrast to that these parameters are frozen in other case, when downstream task is performed. The best results are marked with bold fonts. The result marked as bold is shown a statistically significant improvement to baselines at the 0.01 level using Wilcoxon signed-rank test.

As shown in Table 4, firstly we can find that the pre-trained model w/o LAT using multilingual data does not perform much better than directly modeling on non-native data, which is evidenced by that the recall drops a lot. We think the reason is similar to that of phone recognition that excessively introducing learner's native patterns does not benefit discovering the non-native patterns.

By adding language adversarial training, the recall is improved by a large margin than pre-trained model w/o LAT, and outperforms the baseline. Implicit approach Implicit LAT can help find more patterns to improve the recall but the precision decreases a lot. It indicates that the non-native patterns found by the simple task of cross-language contrastive learning may be not accurate enough since the in-language information is missing.

We also noticed that Deep explicit LAT is better than Shallow LAT, which is consistent with the previous results of phone recognition. A

mong different schemes for language adversarial training, Deep explicit LAT w/ fine-tune achieves the best performance and it outperforms the baseline and other schemes of language adversarial.

It also can be found that by fine-tuning the parameters in the Deep explicit LAT, the performance can be further improved.

2.5 Summary

In this chapter, we propose an unsupervised approach to learn representations from a large amount of Chinese and Japanese raw speech data to non-native acoustic modeling

for mispronunciation verification. In our model, an unsupervised model is employed to learn phonetic structures from Chinese speech. Meanwhile, language adversarial learning is introduced using Japanese speech to align the feature distribution between two languages. The experimental results present that for the non-native phone-level speech recognition and mispronunciation verification tasks (1) the knowledge learned from two native languages speech with the proposed unsupervised framework are useful for non-native acoustic modeling of phone recognition and mispronunciation verification.

Chapter 3

Formant Augmented Language Adversarial Representation Learning for Non-Native Acoustic Modeling of Mispronunciation Verification

This chapter presents the language representation learning for mispronunciation verification. The work presented in this chapter has been published in INTERSPEECH 2020 (Yang et al., 2020) and Neural Networks Journal (Yang, Fu, Zhang, & Shinozaki, 2021).

3.1 Introduction

In our previous work, we proposed a self-supervised pre-training approach for non-native mispronunciation verification, designed to address the data sparsity problem commonly encountered in non-native pronunciation tasks. By leveraging knowledge learned from large-scale speech data in two native languages, we hypothesize that this approach can effectively enhance the model's ability to detect and evaluate mispronunciations in non-native speech. Our methodology involves utilizing a large corpus of unlabeled raw speech from the target language to train the model. This data allows the model to capture key phonetic properties by predicting observations within the target language speech. To facilitate this process, we introduce the sinc filter, a technique designed to extract formant-like features from speech signals. After the pre-training phase, the model undergoes language-adversarial training using speech data from the learner's native language. This step enables the model to identify patterns and relationships between the target language and the learner's native language. These patterns are critical, as non-native pronunciation is often influenced by the learner's native language, a phenomenon known as L1 transfer (Iverson et al., 2003). L1 transfer typically manifests as subtle deviations from canonical pronunciations in the target language.
In this chapter, building on the previously proposed self-supervised pre-training frame-

work, we introduce a formant-augmented language-adversarial representation learning approach for non-native mispronunciation verification. One of the most critical components of current waveform-based, convolutional-layer encoders is the first convolutional layer. This layer must handle high-dimensional inputs and is particularly prone to challenges such as the vanishing gradient problem, especially in very deep architectures. The filters learned in this initial layer often exhibit noisy and inconsistent multi-band shapes, a phenomenon further exacerbated by limited training data. While these filters may be effective for the network, they are not always intuitively meaningful to humans and may not efficiently capture the speech signal's essential characteristics. To enhance both the interpretability and efficiency of the filters in the input layer, we propose replacing the first convolutional layer with a sinc filter to extract formant-like features. Formants are closely related to certain types of mispronunciations, particularly in terms of places and manners of articulation (Wu & Lin, 1989). This information is not only helpful for detecting pronunciation errors but also for providing detailed, instructive feedback to guide learners in correcting their mispronunciations.

3.2 Formant Augmented Language Adversarial Representation Learning

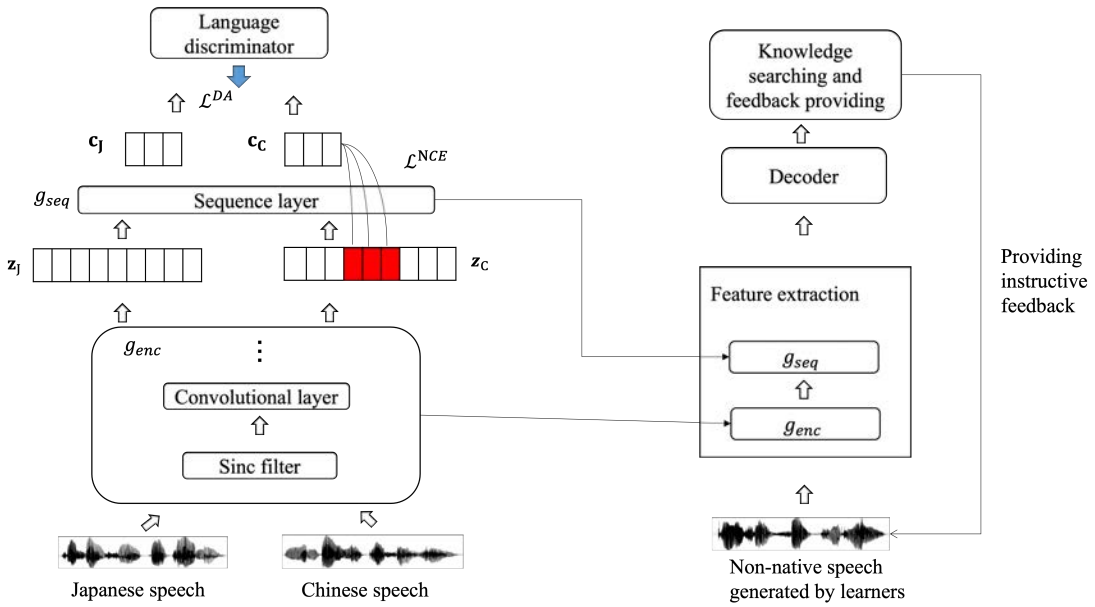

Figure 7: A demonstration of our proposed formant augmented language adversarial representation learning framework.

Our goal is to obtain the representations of the knowledge from the raw speech data from two native languages for non-native acoustic modeling without human supervision. Figure 7 demonstrates our proposed formant augmented language adversarial representation learning framework. In the pre-training stage shown on the left, the model accepts the target language (Chinese in this work) data as input to capture phonetic structure in an unsupervised manner. Meanwhile, the learner's native language (Japanese) speech is fed into the model to align the feature distribution between two languages. Note that Japanese speech data do not participate in calculating the loss $\mathcal{L}^{NCE}$ of the model and only take part in the adversarial loss $\mathcal{L}^{DA}$ of language discriminator. The solid arrow means that the loss is fed through a gradient reversal layer to confuse the language discriminator. After pre-training, the pre-trained model is incorporated in the downstream mispronunciation verification framework.

3.2.1 Encoder

An overview of the model is illustrated in the solid block in the left part of Figure 7. In detail, let $\mathbf{x} = \{x_1, x_2, ..., x_l\}, x_i \in \mathbb{R}$, denotes an raw speech signal in L discrete time steps where x_i is the acoustic amplitude at time i. First, an encoder g_{enc} encodes the signal into embedding vector representations $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_t\}, \mathbf{z}_i \in \mathbb{R}^{d_z}$, where d_z is the dimension of the hidden representation.

In the encoder g_{enc} of the previous models, a standard convolutional layer performs a set of time-domain convolutions between the input waveform and some Finite Impulse Response (FIR) filters, which can be defined as follows:

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l] \tag{15}$$

where $x[n]$ is a chunk of the speech signal, $h[n]$ is the filter of the length M, and $y[n]$ is the output of the filter. In this case, all the elements of $h[\cdot]$ are learnable parameters.

The main difference in this work is that sinc filter is employed in the first layer, which is reported to be able to extract formant-like features (Ravanelli & Bengio, 2018b). In previous research, formant is found related to the placements and manners of articulation, e.g., the first formant (F1) and the second formant (F2) are related to the tongue

position, and the third formant (F3) is related to the shape of lip. Standard features, for instance, smooth the speech spectrum, possibly hindering the extraction of these characteristics (Ravanelli & Bengio, 2018a).

The operation of sinc filter performs a predefined function g that depends on few learnable parameters θ:

$$y[n] = x[n] * g[n, \theta] \tag{16}$$

g is defined as a filter-bank composed of rectangular bandpass filters inspired by standard filtering in digital signal processing.

The magnitude of a generic bandpass filter can be written as the difference between two low-pass filters in the frequency domain:

$$G[f, f_1, f_2] = rect(\frac{f}{2f_2}) - rect(\frac{f}{2f_1}) \tag{17}$$

where f_1 and f_2 are the learned low and high cutoff frequencies, and $rect(\cdot)$ is the rectangular function in the magnitude frequency domain.

The reference function g in time domain can be denoted as:

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{18}$$

where sinc function is defined as $sinc(x) = sin(x)/x$.

Subsequently, the output of filters is sent to subsequent convolutional layers to generate embedding vector representations $\mathbf{z}$, which denotes the hidden representations of speech signal.

The brief process of the encoder g_{enc} can be written:

$$\mathbf{z} = g_{enc}(x_1, x_2, ..., x_l) \tag{19}$$

Then a sequence model g_{seq} summarizes the past information of vector and produces corresponding context-aware representations, which can be denoted as $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t\}, \mathbf{c}_i \in$

$\mathbb{R}^{d_c}$, i.e.,

$$\mathbf{c} = g_{seq}(\mathbf{z}), \tag{20}$$

The optimization is performed by minimizing the InfoNCE (Oord et al., 2018), which is a loss function based on noise contrastive estimation as the lower-bound to the mutual information between context aware embedding $\mathbf{c}_t$ and future latent representations $\mathbf{z}_{t+k}$ for $k \in \{1, ..., K\}$. Given a set $Z = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ which contains one positive sample from $p(\mathbf{z}_{t+k}|\mathbf{c}_t)$ and $N-1$ negative samples from "noise" distribution $p(\mathbf{z})$. The calculation of InfoNCE can be described in Figure. The InfoNCE loss function for each step t can be donated as follows:

$$\mathcal{L}_{tk}^{NCE} = -\mathbb{E}\left[\log \frac{f_k(\mathbf{c}_t, \mathbf{z}_{t+k})}{\frac{1}{N}\sum_{\tilde{\mathbf{z}} \in Z} f_k(\mathbf{c}_t, \tilde{\mathbf{z}})}\right] \tag{21}$$

where $f_k(\mathbf{c}_t, \mathbf{z}_{t+k})$ is a scoring function that can be a lob-bilinear model:

$$f_k(\mathbf{c}_t, \mathbf{z}_{t+k}) = \exp(\mathbf{c}_t^T \mathbf{W}_k \mathbf{z}_{t+k})) \tag{22}$$

where $\mathbf{W}_k$ is the parameters in each model for each k. The total loss to be minimized is a sum of the InfoNCE loss for each step:

$$\mathcal{L}^{NCE} = \sum_t \sum_k \mathcal{L}_{tk}^{NCE} \tag{23}$$

in which negative samples are sampled uniformly from representations in the same speech signal $\mathbf{z}$ in the previous model.

3.2.2 Language adversarial training

The language adversarial training is similar to the previous models. We also perform explicit and implicit adversarial training for a comparison.

3.3 Experiments and Results

Table 5: The detail of non-native dataset.

Text	301
Speakers	7
Number of utterances	1899
Number of phonemes	26431
Average length per utterance	14
Number of annotators	6
Number of annotators per utterance	2

3.3.1 Datasets

Native corpora

AISHELL corpus is employed as the Mandarin Chinese source, which is an open-source Mandarin Chinese speech corpus (Bu et al., 2017). And Corpus of Spontaneous Japanese (CSJ) is used as the Japanese source, which is a database containing a large collection of Japanese spoken language data (Maekawa et al., 2004). We randomly choose 300 hours of data from two corpora (150h from Chinese, 150h from Japanese) above as our training set for pre-training.

Non-native corpus

BLCU inter-Chinese speech corpus, which is collected for language learners who learn Mandarin Chinese as their second language (Cao et al., 2010), is employed as our non-native dataset for mispronunciation verification. This work focuses on the Japanese part as shown in Table 5. It contains speech from 17 Japanese speakers. Each speaker generate 301 utterances in mandarin and totally 4,631 utterances involving 64,190 phonemes are included. Recordings are made in the sound-proofing speech lab with the sampling rate of 16kHz then encoded in 16-bit pulse-code modulation (PCM). All ground-truth labels in this corpus are annotated according to the discussion of well-trained phoneticians. Around 80% of this corpus is used as the training set, 10% is used as the developing set and the rest for testing. There is no overlap of speakers between the training and testing set and leave-one-out cross-validation (Browne, 2000) is adopted.

3.3.2 Experimental setup

Non-native only

We first establish the mispronunciation verification framework using only non-native speech data as one of our baselines. We establish several models explored in previous researches (Gao et al., 2015; Lin et al., 2020; Yang et al., 2017) including

- deep feed-forward neural network (DNN) with five layers where each layer has 550 units,
- convolutional neural network (CNN) with three convolutional layers where the size of layers are [128, 256, 512] and size of filters are [3, 3, 3] sequentially,
- bi-directional recurrent neural network with gated recurrent units (GRU-RNN) with one layer of 550 units (Cho et al., 2014) for acoustic modeling.

The 40-dimensional surface feature Mel-frequency cepstral coefficients (MFCCs) extracted from non-native data through 25ms windows with a 10ms frameshift is employed as the input feature. All the feature are applied cepstral mean and variance normalization (CMVN) (Viikki & Laurila, 1998) before training. Batch normalization (Ioffe & Szegedy, 2015) and a dropout (Srivastava et al., 2014) of 0.2 are employed following each layer. RmsProp (Tieleman & Hinton, 2012) is employed as the optimizer with a batch size of 16.

Pre-trained part

Pretrained model w/o LAT with sinc. Figure 8 presents one of the baselines for pre-training model without LAT with sinc filter using multilingual data. We mix the data from two languages (Mandarin Chinese and Japanese) into a large multilingual dataset for training this setting. The encoder contains five 1-dimensional convolutional layers with a 160 down-sampling factor thus there is a feature vector for every 10ms of speech, which keeps consistent with the rate of phoneme sequence labels obtained with Kaldi. We replace the first convolutional layer with sinc filter. For convolutional layers, the size of filters are [8, 4, 4, 4], the strides are [4, 2, 2, 2] and the paddings are [2, 1, 1, 1]. 512 hidden units of each layer are with ReLU activation. Batch normalization is

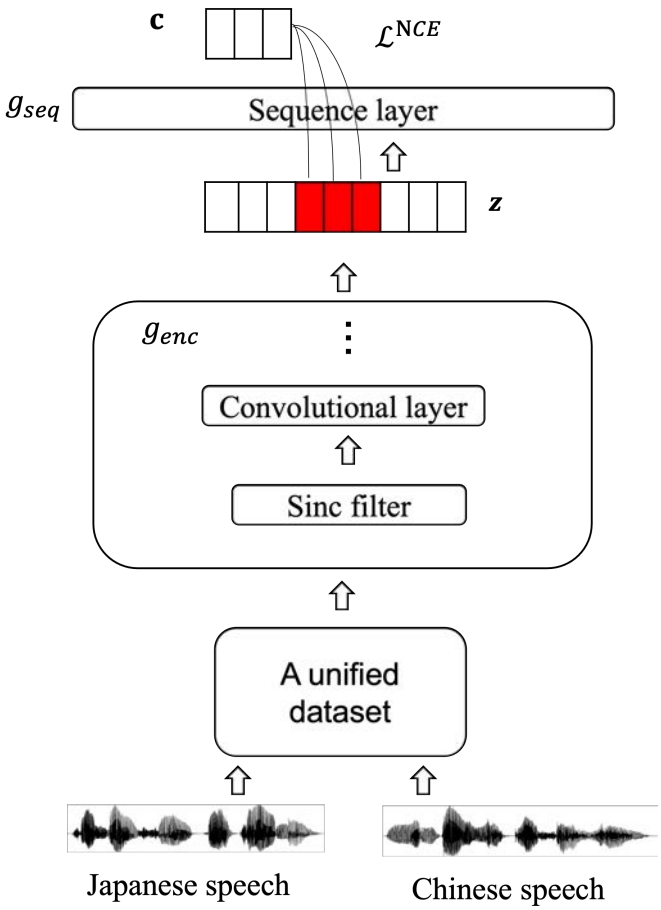

Figure 8: A demonstration of pre-trained model w/o LAT with sinc filter.

employed following each convolutional layer. A recurrent neural network with gated recurrent units (GRU-RNN) with 256-dimensional hidden state is employed as the sequence model. The output of GRU at every timestep is used as the context $\mathbf{c}$ to predict 12 timesteps in the future. Adam optimizer (Kingma & Ba, 2014) with a learning rate of 2e-4 is used to train the model with a minibatch whose size is 8. In each training iteration, a segment containing 20480 data points (around 1.28s) is randomly selected from the speech for every utterance.

Implicit LAT with sinc. For implicit language adversarial training, there is no additional language discriminator. Figure 9 presents the implicit LAT with sinc filter. We replace the first convolutional layer with sinc filter. The main difference is that, when calculating InfoNCE, the negative samples are selected from the learner's native language (Japanese in this work), which reflects the idea of language adversarial.

We explore two explicit language adversarial training approaches shown as follows:

Shallow explicit LAT with sinc. In this approach, We replace the first convolutional

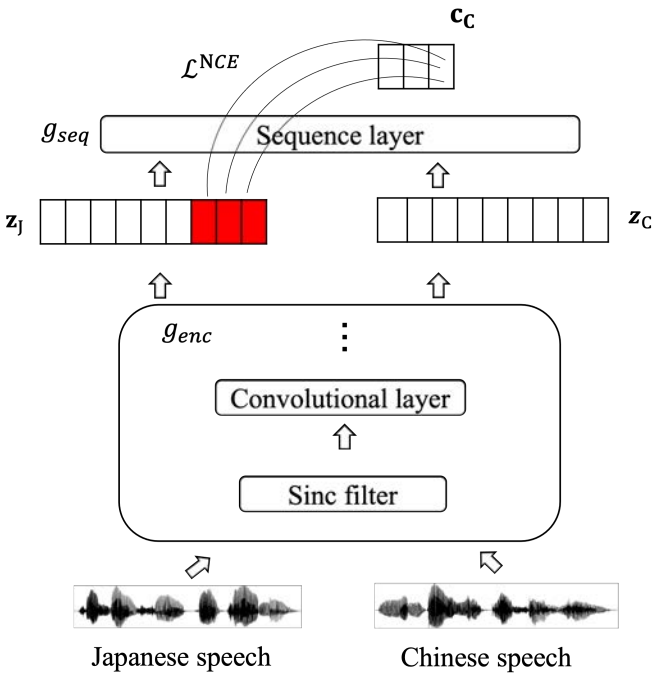

Figure 9: A demonstration of implicit LAT with sinc filter.

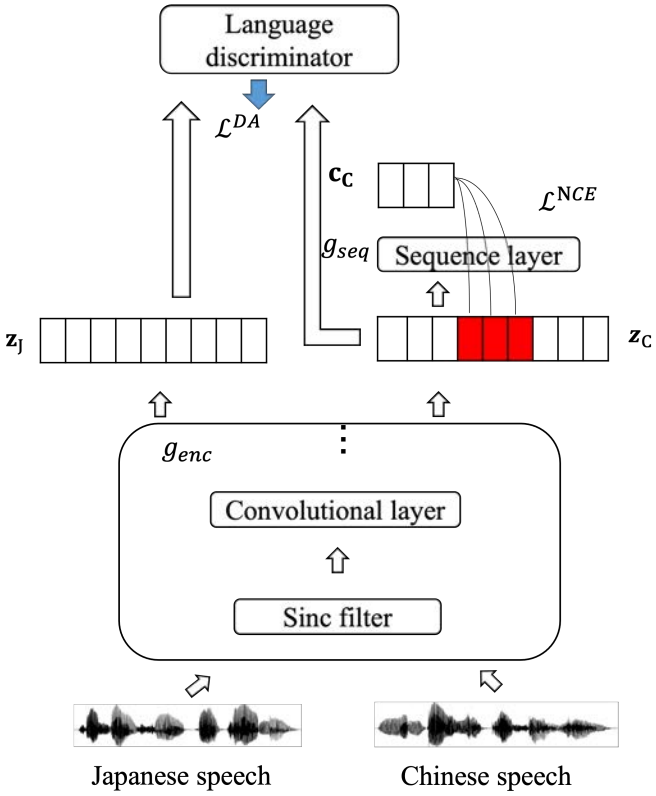

Figure 10: A demonstration of shallow explicit LAT with sinc filter.

layer with sinc filter and take the output of the last convolutional layer as the input into the language discriminator. Figure 10 demonstrates the shallow explicit LAT with sinc

filter. The language discriminator contains two layers with hidden units whose size is [64, 2], and the final layer output the one-hot representations of two languages. λ in Eq (9) are set to 0.5.

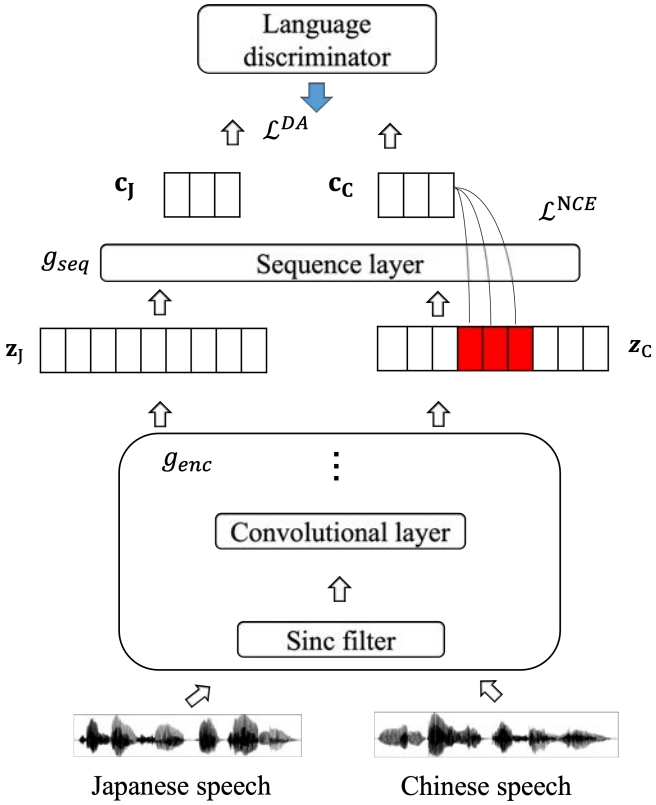

Figure 11: A demonstration of deep explicit LAT with sinc filter.

Deep explicit LAT. The main difference from deep explicit lat is that the output of the sequence model is fed into the language discriminator. Figure 11 demonstrates the deep explicit LAT with sinc filter. The configuration of language discriminator is similar to that of shallow explicit LAT. We replace the first convolutional layer with sinc filter.

Downstream non-native part Then we employ our pre-trained model in the mispronunciation verification framework in which the pre-trained model serve as a feature extractor. The output, 256-dimensional vector representation, of the pre-trained model is directly sent to the downstream decoder (Note that with pre-trained models, only the vector representations are employed as features). The mispronunciation verification framework is a hybrid neural network system. The senone labels (tied HMM states) are first obtained by a Gaussian mixture model-hidden markov model (GMM-HMM), then these labels and the corresponding aligned frames are used for training. The same

align information is employed by all schemes. The decoder use one layer of bi-GRU with 550 units. Batch normalization and a dropout of 0.2 are performed following each layer. RmsProp is employed as the optimizer with a batch size of 16. The configurations are the same for all of the pre-training approaches for a fair comparison. Whether the parameters of the pre-trained model is trainable is explored in the experiments.

3.3.3 Evaluation metrics

3.3.3.1 Metrics for non-native phone recognition

The CAPT system must first be capable of recognizing non-native speech in high performance. In this work, our primary focus is on evaluating the performance at the phone level, encompassing both standard pronunciations and predefined mispronunciations. A key motivation for employing phone error rate (PER) as a metric is its ability to assess the system's performance in recognizing all phones. It is important to note that mispronunciations constitute only a small fraction of all phones and are inherently a subset of the total phone set. This imbalance underscores the criticality of accurate recognition, as any errors in identifying correct pronunciations as mispronunciations could lead the model to generate misleading feedback for learners. Such inaccuracies would compromise the system's reliability and tamp down learners' enthusiasm. Therefore, achieving robust overall performance across all phonemes is essential to ensure the CAPT system provides consistent and constructive feedback.
The PER can be defined as follow:

$$PER = \frac{S + D + I}{N} \tag{24}$$

where N is the total number of all the phones. S, D and I are the numbers of substitution, deletion and insertion respectively.

3.3.3.2 Metrics for non-native mispronunciation verification

The recall and precision are employed as the evaluation metrics for mispronunciation verification. The $recall$ measures that, among all of the phones labeled as the errors manually, how many errors are detected by the detection system. The $precision$ mea-

sures that how many mispronunciations detected by the system are truly pronunciation errors. $F1 - score$ is used since we consider that the precision and recall are equally important for the language learners when using CAPT systems. DA (detection accuracy) measures the overall performance of mispronunciation verification. Among all of the 65 kinds of mispronunciations that occurred in the corpus, 16 most common errors were selected for analysis.

$$Recall = \frac{TR}{TR + FA} \tag{25}$$

$$Precision = \frac{TR}{TR + FR} \tag{26}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{27}$$

$$DA = \frac{TA + TR}{TA + TR + FA + FR} \tag{28}$$

Where TR (true rejection) notes the number of phones labeled as the errors by the expert at the same time detected as the errors by the detection system. TA (true acceptance) denotes the number of phone segments that are marked as the correct pronunciation by the system and the ground truth. FR (false rejection) refers to the number of phones recognized as pronunciation errors by the system while the ground truths are correct. FA (false acceptance) are the number of phone segments that are misrecognized as correct while they are actually errors.

3.4 Results

3.4.1 Results of Non-native phone recognition

Table 6: Detection performance of phone recognition for different approaches.

Model	PER
Non-native only	
DNN	12.25%
CNN	12.09%
GRU	11.22%
Previous proposed pre-trained SSL approaches	
Pre-trained model w/o LAT	10.78%
Implicit LAT	10.37%
Shallow explicit LAT	11.03%
Deep explicit LAT	9.85%
Previous proposed pre-trained SSL approaches w/ sinc filter	
Pre-trained model w/o LAT w/ sinc	10.43%
Implicit LAT w/ sinc	10.09%
Shallow explicit LAT w/ sinc	10.54%
Deep explicit LAT w/ sinc	9.9%
Previous proposed pre-trained SSL approaches w/ sinc filter w/ fine-tune	
Pre-trained model w/o LAT w/ sinc fine-tune	10.33%
Implicit LAT w/ sinc fine-tune	9.87%
Shallow explicit LAT w/ sinc fine-tune	10.11%
Deep explicit LAT w/ sinc fine-tune	**9.59%**

To evaluate the performance, the system should recognize phones besides that not defined as mispronunciations first. Table 6 demonstrates the detection performance of phone recognition with different approaches. "Non-native only" denotes that the model we directly train on only non-native data. "Implicit LAT" is the implicit language adversarial training with negative sampling from target language. "Shallow explicit LAT" denotes we use the output of encoders as the input fed into language discriminator and "Deep explicit LAT" means we use the output of sequence model for language discriminator. "fine-tune" denotes that the parameters of the pre-trained model are train-

able, in contrast to that these parameters are frozen in other case, when downstream task is performed. The best results are marked with bold fonts. The result marked as bold is shown a statistically significant improvement to baselines at the 0.01 level using Wilcoxon signed-rank test.

w/ sinc filter vs. w/o sinc filter

It also can be found that introducing sinc filter can further improve the performance of most of the models. By introducing sinc filter, the performance of pre-trianed model w/o LAT with multilingual data is improved from 10.78% to 10.43%, the performance of implicit LAT is improved from 10.37% to 10.09%, the performance of shallow explicit LAT is improved from 11.03% to 10.11%. The performance of deep explicit LAT decrease a little from 9.85% to 9.9%. We wonder it may be due to the difference between the data of pre-training and downstream inferring. By fine-tuning the parameters in it, the deep explicit LAT w/ sinc fine-tune achieves the best results of 9.59%.

Table 7: Detection performance of mispronunciation verification for different approaches.

Model	Recall	Precision	F1 score	DA
Non-native only				
DNN	38.97%	48.14%	43.07	82.04%
CNN	40.35%	48.88%	44.2	82.55%
GRU	40.12%	54.92%	46.37	84.66%
Proposed pre-trained approaches				
Pre-trained model w/o LAT	34.73%	54.2%	42.33	84.29%
Implicit LAT	41.92%	50.72%	45.90	84.88%
Shallow explicit LAT	37.72%	51.79%	43.64	84.79%
Deep explicit LAT	44.91%	58.14%	50.68	86.44%
Previous proposed pre-trained SSL approaches w/ sinc filter				
Pre-trained model w/o LAT w/ sinc	37.13%	58.49%	45.42	84.78%
Implicit LAT w/ sinc	41.23%	52.99%	46.37	84.97%
Shallow explicit LAT w/ sinc	38.10%	52.21%	44.05	84.89%
Deep explicit LAT w/ sinc	44.14%	60.22%	50.94	85.93%
Previous proposed pre-trained SSL approaches w/ sinc filter w/ fine-tune				
Pre-trained model w/o LAT w/ sinc fine-tune	39.25%	58.71%	47.04	84.85%
Implicit LAT w/ sinc fine-tune	42.77%	54.56%	47.95	85.33%
Shallow explicit LAT w/ sinc fine-tune	40.27%	53.88%	46.09	84.98%
Deep explicit LAT w/ sinc fine-tune	**45.33%**	**60.57%**	**51.72**	**86.98%**

3.4.2 Results of non-native mispronunciation verification

Table 7 demonstrates the detection performance of non-native mispronunciation verification with different approaches. “Implicit LAT” is the implicit language adversarial training with negative sampling from target language. “Shallow explicit LAT” denotes we use the output of encoders as the input fed into language discriminator and “Deep

explicit LAT" means we use the output of sequence model for language discriminator. "fine-tune" denotes that the parameters of the pre-trained model are trainable, in contrast to that these parameters are frozen in other case, when downstream task is performed. The best results are marked with bold fonts. The result marked as bold is shown a statistically significant improvement to baselines at the 0.01 level using Wilcoxon signed-rank test.

w/ sinc filter vs. w/o sinc filter

As shown in Table 7, firstly we can find that by introducing sinc filter, the performance is improved for most of the settings. Specifically, the performance about precision is improved to a relatively large margin. The precision of pre-trained model w/o LAT is improved from 54.2% to 58.49%, the precision of implicit LAT is improved from 50.72% to 52.99%, the precision of shallow explicit LAT is improved from 51.79% to 52.21%, the precision of deep explicit LAT is improved from 58.14% to 60.22%. By fine-tune the parameters in it, the deep explicit LAT w/ sinc fine-tune achieves the best results, which has recall of 45.33%, precision of 60.57%, F1 socre of 51.72, and DA of 86.98%.

A comprehensive analysis

To make a detailed analysis, we group the employed mispronunciations into four groups and divide the results , as shown in Figure 12. Four groups, i.e.

- The shape of lip is rounding or spreading: sounds with spreading lips have problems of rounding tendency or sounds with the rounding lips have problems of spreading tendency
- The position of the tongue is advancing and backing: the tongue position of phonemes is a little advance or back
- The aspiration or constriction is sufficient or not
- Laminalizing: some balade-palatal phonemes are pronounced like Japanese lamina-alveolar

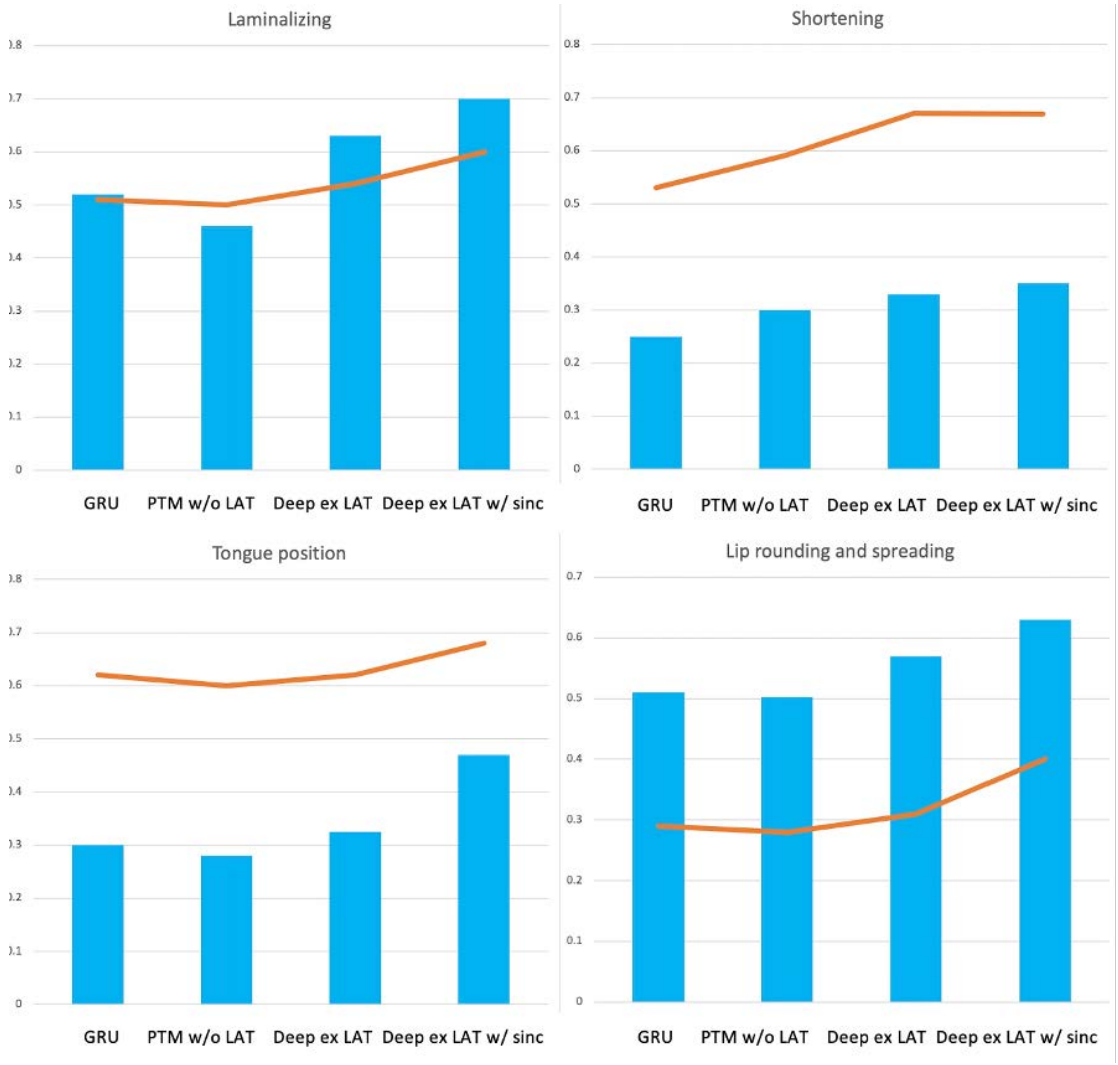

Figure 12: Results for four groups of mispronunciations. Column is $Recall$ and line is $Precision$.

As shown in Figure 12, overall we notice that our model improves the recall for all four groups. It also can be found that the recall for detecting errors about laminalizing, and the shape of lip is improved by a large margin by introducing language adversarial training, which indicates that the patterns of these three kinds of errors are captured by our approaches. By introducing sinc filters, the recall and precision of tongue position and the shape of lip are improved a lot. It meets our presumption and is consistent with previous research (Wu & Lin, 1989) about phonetic indicating that the first formant (F1) and the second formant (F2) formant are sensitive to the tongue position (high and low to F1, front and rear to F2), and the third formant (F3) is related to the shape of lip (round or spread).

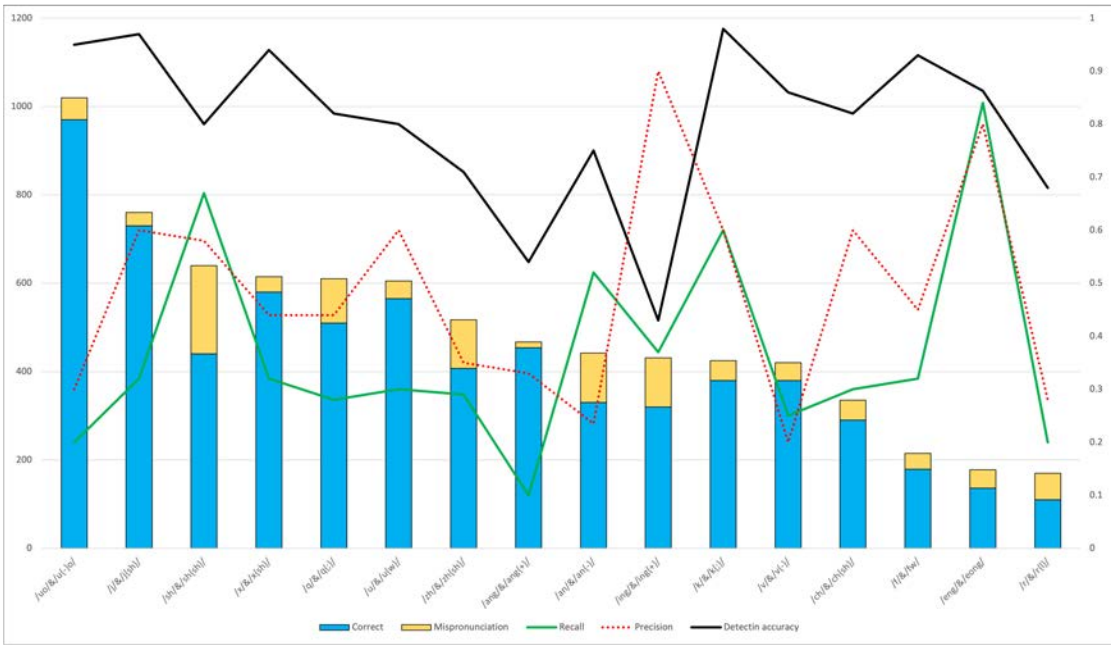

Figure 13: The detailed results of used 16 most frequent mispronunciations using non-native data only with GRU model.

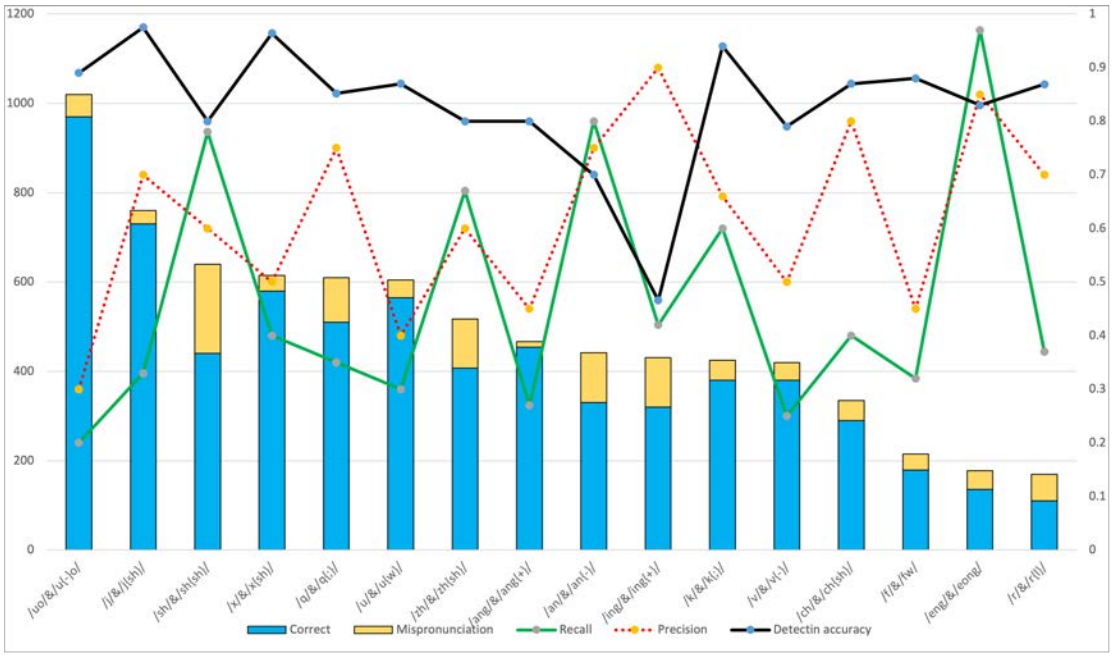

Figure 14: The detailed results of used 16 most frequent mispronunciations using Deep explicit LAT with sinc filter with refreezing parameters.

A demonstration of the performance about each mispronunciation.

For detailed information, we demonstrate the performance of the proposed unsupervised framework as shown in Figure 13, Figure 14. From the figures, it can be easily found the performance for those errors that have lower numbers using non-native only

is low, for example, the detection accuracy of /r/&/rl/ is below 70% and the recall and precision is below 30%. It is difficult to capture the patterns of these kinds of errors using a limited dataset. It is because the skewed distribution of class instances forces the model to be biased to majority classes (He & Shen, 2007) thus it is a little possibility for the model to find these error patterns. Our approach alleviate this problem to a certain degree, which is reflected in that the performance for the errors of lower amount, e.g., /f/&/fw/, /eng/&/eong/ and /r/&/rl/, are improved. Another phenomenon can be found the proportion of mispronunciations and the correct ones are generally imbalanced, such as /j/&/jsh/, /x/&/xsh/, /u/&/uw/, /ang/&/ang+/, /k/&/k;/ and /v/&/v-/, which may lead poor performance when setting up models using non-native directly. Our proposed model is shown its ability to handle this problem as well. These kinds of non-native patterns can be found in the process of pre-training stage with a large amount of learner's target and native languages and then employed in the downstream task. The figures indicate that the performance of /u/&/uw/, /ang/&/ang+/ and /v/&/v-/ are improved to various extents.

3.5 Summary

In this chapter, we propose an formant augmented self-supervised learning approach to learn representations from a large amount of Chinese and Japanese raw speech data to non-native acoustic modeling for mispronunciation verification. On the basis of our previously proposed language adversarial framework, we replace the first convolutional layer with sinc filters to introduce formant-like feature, which is considered relevant to some kinds of mispronunciations from the respect of placements and manners of articulation. his information is useful not only for detecting pronunciation errors but also to be able to provide detailed and instructive feedback to guide the learners to correct their pronunciation errors. The experimental results present that, for the non-native phone-level speech recognition and mispronunciation verification tasks, formant-like feature can be incorporated by introducing sinc filter to further improve the precision for mispronunciation verification.

Chapter 4

Self-supervised learning with multi-target contrastive coding for Non-Native Acoustic Modeling of Mispronunciation Verification

This chapter presents the self-supervised learning with multi-target contrastive coding for non-native mispronunciation verification. The work presented in this chapter has been published in INTERSPEECH 2022 (Yang et al., 2022).

4.1 Introduction

The flexibility in time and space, along with low costs, has led to increasing attention on computer-aided language learning (CALL) systems in recent years. As a core component of CALL systems, computer-aided pronunciation training (CAPT) functions as a virtual pronunciation teacher. It processes and analyzes language learners' non-native speech, providing an assessment of pronunciation quality, known as pronunciation assessment (Hu et al., 2013b; Witt & Young, 2000; Zheng et al., 2007). More specifically, CAPT should be capable of accurately detecting pronunciation errors, diagnosing the types and locations of these errors in learners' utterances, and offering instructive feedback on how to correct them in future learning, which is referred to as mispronunciation verification (Harrison et al., 2009; Hu et al., 2015; Y.-B. Wang & Lee, 2012; Wei et al., 2009).

Most mainstream mispronunciation verification systems are based on state-of-the-art speech recognition technologies, with non-native acoustic modeling playing a crucial role. Given the numerous variations in non-native speech, such as non-idiomatic pronunciation placements, manner distortions, and disfluencies, it is relatively straightforward to develop an acoustic model using the learner's non-native data. With the recent advancements in deep learning, mispronunciation verification has evolved, benefiting from machine learning techniques in several studies (Duan et al., 2014; Gao et al., 2015;

Yang et al., 2017). However, the approaches in these studies, which rely on supervised learning, require large amounts of training data. Collecting such data is challenging, and manual labeling of large-scale datasets, when available, is time-consuming, as it depends on the expertise of professional phoneticians in speech perception. Consequently, the data sparsity problem remains a significant challenge in non-native acoustic modeling for mispronunciation verification.

In recent years, several studies have explored a variety of approaches for learning audio representations in an unsupervised manner (Hyvarinen & Morioka, 2016; Oord et al., 2018; Rivière et al., 2020; Schneider et al., 2019). (Oord et al., 2018) introduced contrastive predictive coding (CPC), a method that captures high-level properties of speech, such as phonetic structures and speaker characteristics, from raw speech without human supervision. The core idea is to train an encoder that makes long-term predictions about future observations while preserving certain structures or properties of the input. These representations can then be applied to downstream tasks, such as automatic speech recognition (ASR) and speaker verification (Oord et al., 2018). Building on this scalable and cost-effective method, (Yang, Fu, Zhang, & Shinozaki, 2021; Yang et al., 2020) conducted preliminary studies on non-native acoustic modeling using an unsupervised approach based on CPC and adversarial training to capture non-native phonetic patterns from raw speech in two native languages, achieving promising results. However, speech signals are not only high-dimensional and variable-length sequences, but they also involve complex hierarchical structures that are challenging to capture using a single self-supervised task framework (Pascual et al., 2019).

In this chapter, we explore to enrich the representations learned by self-supervised training for non-native acoustic modeling, mainly including two schemes:

- we propose multi-target contrastive coding. Several researches have indicated that the predictive coding-based approaches contrastive to different targets can capture different information, e.g., the model with negative samples from the same utterance to the input can capture phonetic structure (Oord et al., 2018), and which with the targets from different speakers' utterances can learn speaker-related information (Mirco & Yoshua, 2019). Inspired by it, in this work, our model is designed to make predictions about the observations contrastive to different targets jointly to learn the representations of the discrepancy with respect

to phonetic structures in and across languages, and speakers at the same time.

- an additional term to reconstruct the original speech from the shared components. This term serves as a regularization that leads the intermediate representations learned by the model to be a good abstraction of the input speech. With these two schemes, we expect that the representations learned by our propose approaches from two native languages will be transferable and meaningful for non-native acoustic modeling of phone recognition and mispronunciation verification without human supervision. The performance of our proposed method is evaluated on the Japanese part of BLCU inter-Chinese speech corpus, which is designed for learners from Japan who learn Mandarin Chinese as their second language.

4.2 Self-Supervised Learning with Multi-Target Contrastive Coding

Our goal is to learn the representations of the discrepancy with respect to phonetic structures, speakers, and languages at the same time from the raw speech of two native languages without human supervision, then apply them to non-native mispronunciation verification. As shown in Figure 15, the raw Chinese speech is firstly fed into an encoder to be converted to hidden representations, and the model is trained with these representations by making predictions that are contrastive to different targets, i.e., the same and different utterances in the same language, and the utterance from a different language (Japanese). Then the pre-trained model plays a part of the acoustic modeling of the downstream mispronunciation verification task. In this section, we elaborate on the detail of our approaches. Pre-training phase is in the dashed box on the left, and mispronunciation verification using the features extracted by the pre-trained model is on the right. The blocks marked as red are parts of the negative samples when making predictions at t with $\mathbf{C}_t$. $\hat{\mathbf{z}}_C$ is another utterance in the same mini-batch and $\mathbf{z}_J$ is that from another language. Note that Japanese utterances serve as negative samples only.
In more detail, let $\mathbf{x}^C = \{x_1^C, x_2^C, ..., x_l^C\}$ and $\mathbf{x}^J = \{x_1^J, x_2^J, ..., x_l^J\}, x_i^C, x_i^J \in \mathbb{R}$ denote raw speech signal in L discrete time steps in Chinese and Japanese respectively. An encoder g_{enc} of multiple convolutional layers firstly converts the signals

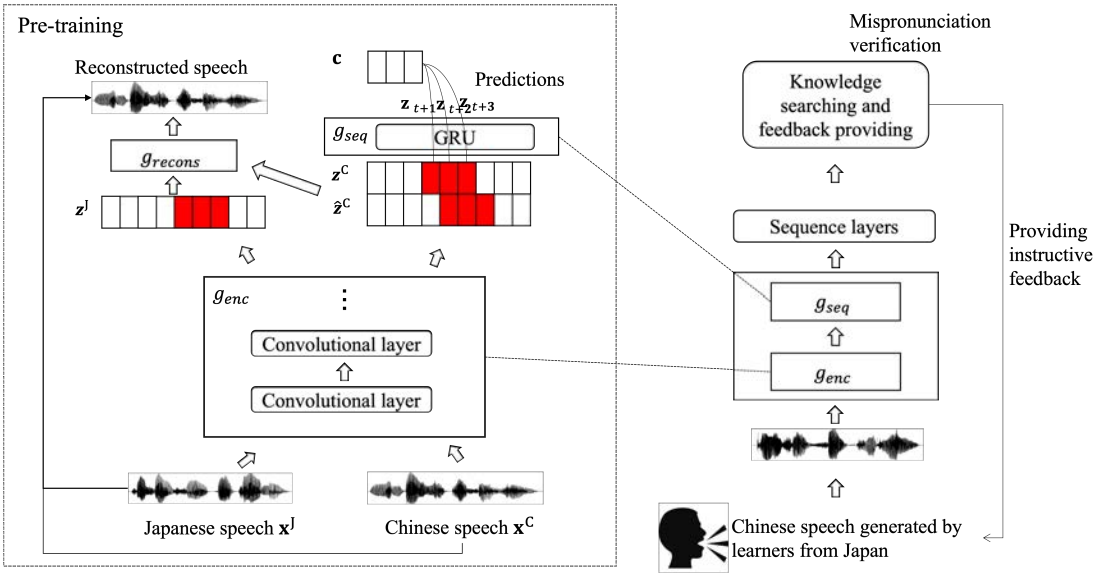

Figure 15: A demonstration of the proposed self-supervised learning with multi-target contrastive coding and mispronunciation verification framework.

into the high-dimensional vector representations $\mathbf{z}^C = \{\mathbf{z}_1^C, \mathbf{z}_2^C, ..., \mathbf{z}_t^C\}$ and $\mathbf{z}^J = \{\mathbf{z}_1^J, \mathbf{z}_2^J, ..., \mathbf{z}_t^J\}, \mathbf{z}_i^C, \mathbf{z}_i^J \in \mathbb{R}^{d_z}$.

$$\mathbf{z}^C = g_{enc}(\mathbf{x}^C) \tag{29}$$

$$\mathbf{z}^J = g_{enc}(\mathbf{x}^J) \tag{30}$$

Then a sequence model g_{seq} generates context aware representations of Chinese speech, which can be defined as $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t\}, \mathbf{c}_i \in \mathbb{R}^{d_c}$.

$$\mathbf{c} = g_{seq}(\mathbf{z}_1^C, \mathbf{z}_2^C, ..., \mathbf{z}_t^C), \tag{31}$$

Meanwhile, a regularization term serves as a constraint to lead the intermediate hidden representations to be an accurate abstraction of the input by reconstructing the speech input through a reconstructor g_{recons}. The regularization works for both languages. The reconstructor is trained to minimize the mean absolute error (L1) based reconstruction loss as follows,

$$\mathcal{L}^{recons} = \sum_{i}^{L} |g_{recons}(\mathbf{z})_i - x_i| \tag{32}$$

The model is trained by optimizing the loss based on InfoNCE (Oord et al., 2018), which is a loss function of noise contrastive estimation as the lower-bound on the mutual information between the context aware embedding $\mathbf{c}_t$ and negative representations $\mathbf{z}_{t+k}$ for $k \in \{1, ..., K\}$. For example of negative sampling from the same utterance, given a set $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ which contains one positive sample from $p(\mathbf{z}_{t+k}|\mathbf{c}_t)$ and $N-1$ negative samples from "noise" distribution $p(\mathbf{z})$. The InfoNCE loss function for each step t can be defined as follows:

$$\mathcal{L}_{tk}^{N} = -\mathbb{E}_{\mathbf{Z}} \left[\log \frac{f_k(\mathbf{c}_t, \mathbf{z}_{t+k})}{\frac{1}{N} \sum_{\tilde{\mathbf{z}} \in \mathbf{Z}} f_k(\mathbf{c}_t, \tilde{\mathbf{z}})} \right] \tag{33}$$

where $f_k(\mathbf{c}_t, \mathbf{z}_{t+k})$ is a scoring function that can be a log-bilinear model:

$$f_k(\mathbf{c}_t, \mathbf{z}_{t+k}) = \exp(\mathbf{c}_t^T \mathbf{W}_k \mathbf{z}_{t+k}) \tag{34}$$

where $\mathbf{W}_k$ are the parameters in each model for each k.

The same procedure is easily made to obtain the loss $\mathcal{L}_{tk}^{batch}$ contrastive to different texts and speakers. Negative samples are drawn from the representations of another utterance $\hat{\mathbf{z}}^C$ in the same mini-batch, which is from the same language but probably different texts or different speakers, or both. As a consequence, the model can learn the representations that embeds properties from different texts like that in (Pascual et al., 2019) or speaker identities from different speakers, or both. Similarly, the loss $\mathcal{L}_{tk}^{lang}$ is obtained with the negative samples from the utterances of a different language (Japanese in this work). We expect it can capture information across different languages encouraging the model to learn the discrepancy between two languages. The speaker's identity information is implicitly included in this process in which they have different language backgrounds. Then the total loss to be minimized is the average of these losses:

$$\mathcal{L} = Avg(\mathcal{L}^{recons} + \sum_{t} \sum_{k} \mathcal{L}_{tk}^{N} + \mathcal{L}_{tk}^{batch} + \mathcal{L}_{tk}^{lang}) \tag{35}$$

4.3 Experiments

Table 8: The detail of non-native dataset.

Text	301
Speakers	7
Number of utterances	1899
Number of phonemes	26431
Average length per utterance	14
Number of annotators	6
Number of annotators per utterance	2

4.3.1 Corpus

AISHELL corpus is employed as the Chinese source, which is an open-source Mandarin Chinese speech corpus (Bu et al., 2017), and Corpus of Spontaneous Japanese (CSJ) is used as the Japanese source, which is a database containing a large collection of Japanese spoken language data (Maekawa et al., 2004). We randomly choose 150 hours of data from the two corpora above as our training set for pre-training.

BLCU inter-Chinese speech corpus, which is collected for language learners who learn Mandarin Chinese as their second language (Cao et al., 2010), is employed as our non-native dataset. As shown in Table 8, it contains conversational speech from 17 Japanese speakers of 4631 utterances involving 64.190 phonemes. All ground-truth labels of mispronunciations in this corpus are labeled by well-trained phoneticians. Around 80% of this corpus is used as the training set, 10% for validation, and the rest for testing. There is no overlap of speakers between the training and testing sets. Leave-one-out cross-validation is performed in this experiment.

4.3.2 Experiment Setup

Non-native only

Non-Native Only. We first establish a mispronunciation verification model using only non-native speech data as one of our baselines. The 40-dimensional surface feature of Mel-frequency cepstral coefficients (MFCCs) extracted from non-native speech through 25ms windows with a 10ms frameshift is employed as the input feature. The feature

is applied with cepstral mean and variance normalization (CMVN) (Viikki & Laurila, 1998) before training. Bi-directional recurrent neural network with gated recurrent units (GRU-RNN) (Cho et al., 2014) is employed for acoustic modeling. The feature is fed into three-layer bi-GRU (Schuster & Paliwal, 1997), where each layer has 550 units. Batch normalization (Ioffe & Szegedy, 2015) and a dropout (Srivastava et al., 2014) of 0.2 are employed following each layer. RmsProp (Tieleman & Hinton, 2012) is employed as the optimizer with a batch size of 16.

Pre-trained part

Pre-trained model with multilingual data. One of the baselines with pre-training approaches is a pre-trained model with multi-lingual data. We mix the data from two languages (Mandarin Chinese and Japanese) into a large dataset for training the model. The encoder contains five 1-dimensional convolutional layers with a 160 down-sampling factor thus there is a feature vector for every 10ms of speech, which keeps consistent with the rate of phoneme sequence labels obtained with Kaldi. For convolutional layers, the size of filters are [10, 8, 4, 4, 4], the strides are [5, 4, 2, 2, 2] and the paddings are [3, 2, 1, 1, 1]. 512 hidden units of each layer are with ReLU activation. Batch normalization is employed following each convolutional layer. A recurrent neural network with gated recurrent units (GRU-RNN) with 256-dimensional hidden state is employed as the sequence model. The output of GRU at every timestep is used as the context $\mathbf{c}$ to predict 12 timesteps in the future. In each training iteration, speech data from two languages are combined. A segment containing 20480 data points (around 1.28s) is randomly selected from the speech for every utterance. Adam optimizer (Kingma & Ba, 2014) with a learning rate of 2e-4 is used to train the model with a mini-batch whose size is 8. In this scheme, all of the negative samples are selected from the same utterance to the input.

MTCC. Figure 16 demonstrate the MTCC. The setup of multi-target contrastive coding (MTCC) is similar to the Pre-trained model with multilingual data. The main difference is that the negative samples are selected from the same utterance to the input, the different utterances from the same mini-batch, and utterances from a different language. Note that, in this manner, the Japanese speech serve as negative samples only.

MTCC with reconstruction regularization term. Figure 17 demonstrate the MTCC

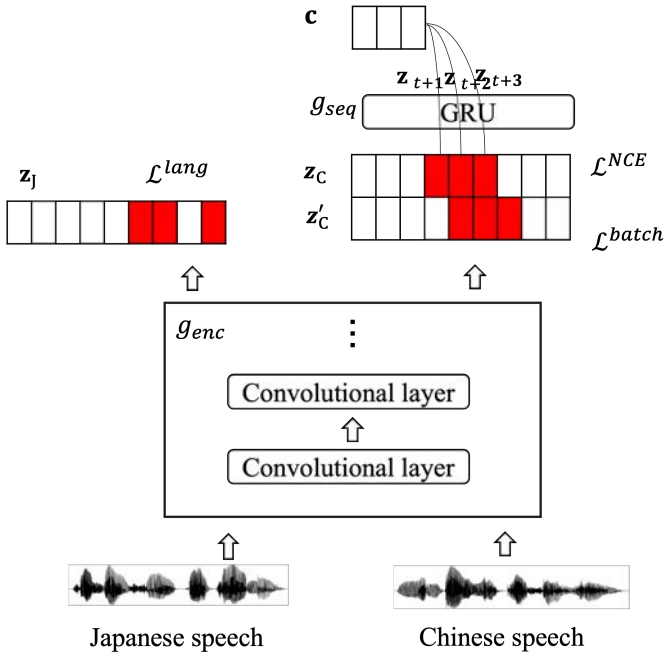

Figure 16: A demonstration of MTCC.

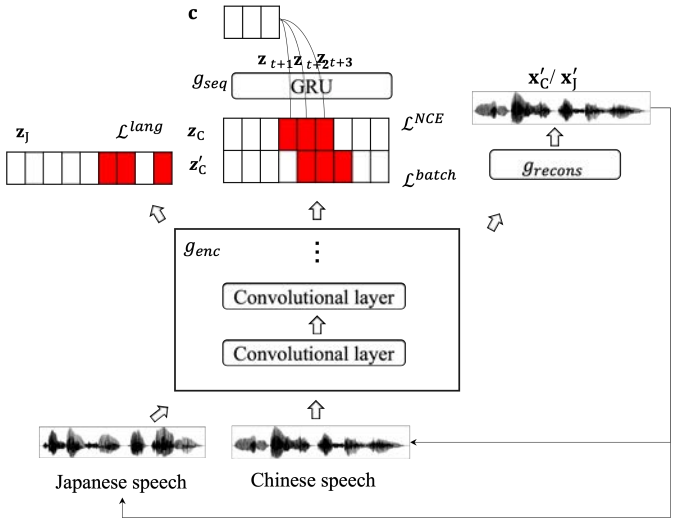

Figure 17: A demonstration of MTCC w/ reconstruction regularization term.

with reconstruction regularization term. The setup of multi-target contrastive coding (MTCC) is similar to the Pre-trained model with multilingual data. The main difference is that a reconstruction regularization term is included to guide the model to learn an accurate representations of the original input.

Then we employ our pre-trained model in the mispronunciation verification task. The mispronunciation verification framework is a hybrid neural network. The senone labels (tied HMM states) are first obtained by a Gaussian mixture model-hidden markov model (GMM-HMM), then these labels and the corresponding aligned frames are used for

training. The same alignment information is employed in all schemes. The decoder is a bi-GRU of one layer with 550 units. Batch normalization and a dropout of 0.2 are performed following each layer. RmsProp is employed as the optimizer with a batch size of 16. The configurations are the same for all of the pre-training approaches for a fair comparison.

4.3.3 Evaluation metrics

4.3.3.1 Metrics for non-native phone recognition

The CAPT system must first be capable of recognizing non-native speech in high performance. In this work, our primary focus is on evaluating the performance at the phone level, encompassing both standard pronunciations and predefined mispronunciations. A key motivation for employing phone error rate (PER) as a metric is its ability to assess the system's performance in recognizing all phones. It is important to note that mispronunciations constitute only a small fraction of all phones and are inherently a subset of the total phone set. This imbalance underscores the criticality of accurate recognition, as any errors in identifying correct pronunciations as mispronunciations could lead the model to generate misleading feedback for learners. Such inaccuracies would compromise the system's reliability and tamp down learners' enthusiasm. Therefore, achieving robust overall performance across all phonemes is essential to ensure the CAPT system provides consistent and constructive feedback.
The PER can be defined as follow:

$$PER = \frac{S + D + I}{N} \tag{36}$$

where N is the total number of all the phones. S, D and I are the numbers of substitution, deletion and insertion respectively.

4.3.3.2 Metrics for non-native mispronunciation verification

The recall and precision are employed as the evaluation metrics for mispronunciation verification. The $recall$ measures that, among all of the phones labeled as the errors manually, how many errors are detected by the detection system. The $precision$ mea-

sures that how many mispronunciations detected by the system are truly pronunciation errors. $F1-score$ is used since we consider that the precision and recall are equally important for the language learners when using CAPT systems. DA (detection accuracy) measures the overall performance of mispronunciation verification. Among all of the 65 kinds of mispronunciations that occurred in the corpus, 16 most common errors were selected for analysis.

$$Recall = \frac{TR}{TR + FA} \tag{37}$$

$$Precision = \frac{TR}{TR + FR} \tag{38}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{39}$$

$$DA = \frac{TA + TR}{TA + TR + FA + FR} \tag{40}$$

Where TR (true rejection) notes the number of phones labeled as the errors by the expert at the same time detected as the errors by the detection system. TA (true acceptance) denotes the number of phone segments that are marked as the correct pronunciation by the system and the ground truth. FR (false rejection) refers to the number of phones recognized as pronunciation errors by the system while the ground truths are correct. FA (false acceptance) are the number of phone segments that are misrecognized as correct while they are actually errors.

4.4 Experiments

4.4.1 Main results

Firstly, we evaluate the performance for non-native phone recognition as mispronunciation verification needs to recognize them first. Phone error rate (PER) is used as the metric. Table 5 presents the phone error rate (PER) for phone recognition with different approaches. Non-native only denotes the acoustic modeling directly on non-native data only. MTCC is the proposed model in this work. FT denotes the parameters of the pre-trained model are fine-tuned at the downstream stage.

As shown in Table 9, it indicates that our proposed model is effective for non-native

phone recognition. Our proposed model improves the PER from 11.94% to 9.84% comparing to modeling using non-native data only. And MTCC outperforms the pre-trained model with multilingual data that is contrastive to the target from the same utterance for all the samples, which obtains the PER of 12.13%. It also can be found that the performance can be improved by introducing reconstruction regularization that guides the hidden vectors to be an accurate abstraction of the input. With it the PER is decreased from 9.84% to 9.42%. We also investigate the performance in the case of unfreezing the parameters of the pre-trained model at the downstream stage to handle the mismatch between different datasets. The results report that the performance can be further improved by tuning works from 9.42% to 9.38%.

Table 9: Phone error rate (PER) for phone recognition with different approaches.

Model	**PER**(%)
Non-native only	11.94
Pre-trained model with multilingual data	12.13
MTCC	9.84
MTCC+recons	9.42
MTCC+recons+FT	**9.38**

Then we evaluate the performance for mispronunciation verification. As shown in Table 10, the overall performance is improved by the MTCC model. Our proposed model improves the Recall from 40.12% to 49.28% comparing to modeling using non-native data only, and the Pre-trained model with multilingual data of a single task from 34.73% to 49.28%. It means that our model can find out more pronunciation errors from all of the phones. It is also can be found that the precision is improved from 54.92% to 60.97% comparing the non-native data only, and from 54.2% to 60.97% comparing to the Pre-trained model with multilingual data, which indicates the number of pronunciation errors detected by our model being truly errors is larger than them. Finally, the best results are obtained by our proposed model with reconstruction regularization. We notice that the performance of the Pre-trained model with multilingual data is not good. When training the Pre-trained model with multilingual data model, speech data from two languages are combined into one dataset and all the negative samples are selected from the same utterance to the input, in which the model performed the predictions within the utterance. While in the scheme of MTCC, the prediction is performed jointly

with multi targets from the same utterance, the utterance of different text, the different speakers, and different languages. This confirms our conjecture that, for the mispronunciation verification task, it would be more effective to enrich the representations with the model of the multi-target self-supervised task than the model of a single self-supervised task.

Table 10: The detection performance for mispronunciation verification.

Model	**Recall**(%)	**Prec**(%)	**F1**	**DA**(%)
Non-native only	40.12	54.92	46.37	84.66
Pre-trained model with multilingual data	34.73	54.2	42.33	84.59
MTCC	49.28	60.97	54.51	86.97
MTCC+recons	51.04	61.43	55.75	87.22
MTCC+recons+FT	**52.31**	**61.87**	**56.69**	**87.59**

4.4.2 Ablation Study

Table 11: Ablation study about the importance of each component in our proposed model.

Model	**PER**(%)	**Recall**(%)	**Precision**(%)
MTCC	9.84	49.28	60.97
w/o $\mathcal{L}^{batch}$	10.33	47.85	58.19
w/o $\mathcal{L}^{lang}$	11.08	42.88	57.34
w/o both	12.13	34.73	54.2

An ablation study is made to investigate the effectiveness of the different components in our proposed model. As shown in Table 11, when removing each of the additional losses, the results decrease to varying degrees. The model degenerates into the Pre-trained model with multilingual data model when removing both of the additional losses. We can also note that the decrease is larger when removing $\mathcal{L}^{lang}$ than that when removing $\mathcal{L}^{batch}$, especially for recall. It indicates that language information plays an important role for this task.

4.4.3 A comprehensive analysis

To make a detailed analysis, we group the employed mispronunciations into four groups and divide the results accordingly. Four groups, i.e.

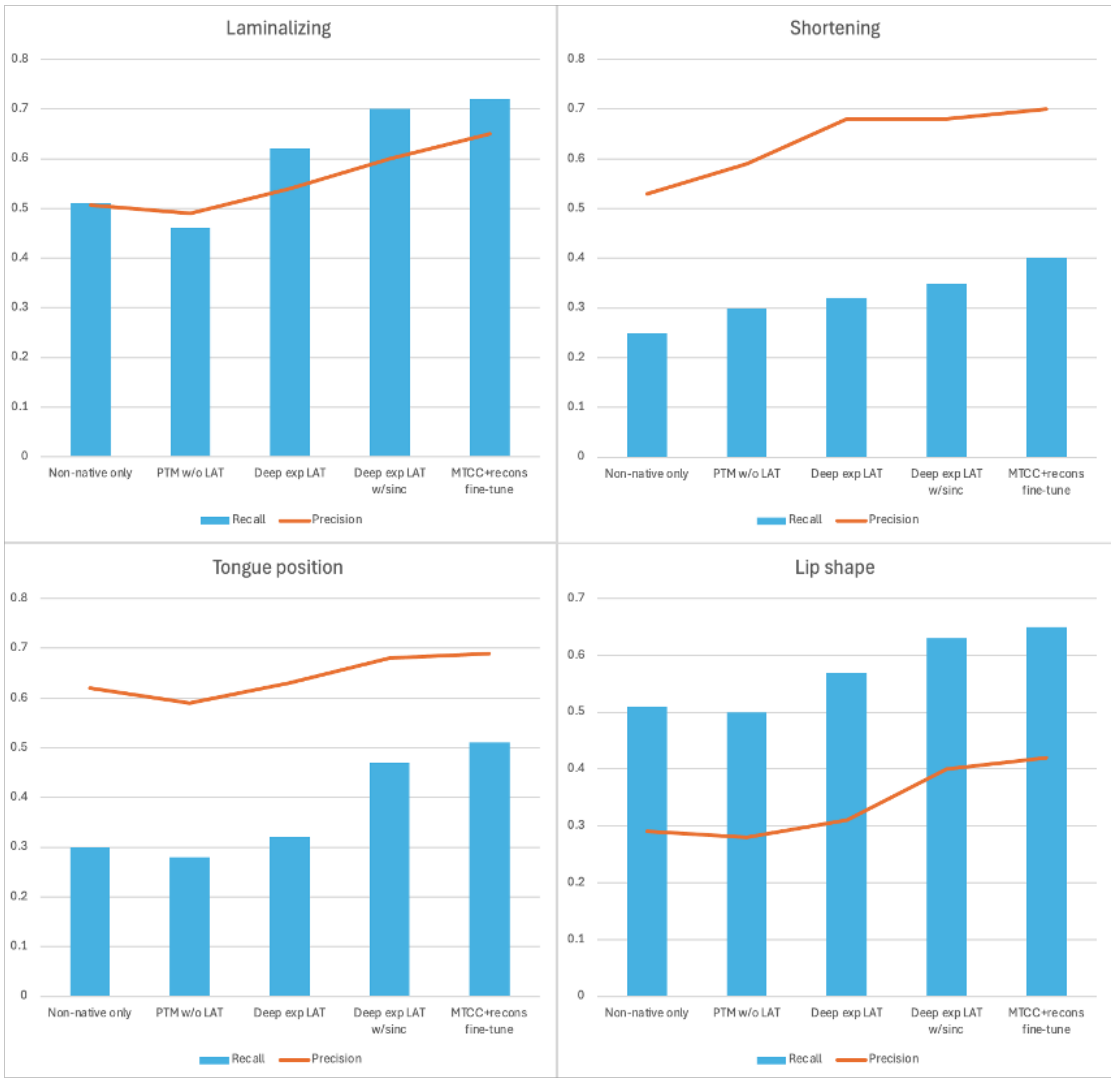

Figure 18: A comprehensive analysis.

- The shape of lip is rounding or spreading: sounds with spreading lips have problems of rounding tendency or sounds with the rounding lips have problems of spreading tendency
- The position of the tongue is advancing and backing: the tongue position of phonemes is a little advance or back
- The aspiration or constriction is sufficient or not
- Laminalizing: some balade-palatal phonemes are pronounced like Japanese lamina-alveolar

As shown in Figure 18, overall we notice that our model improves the recall for all four groups.

4.5 Summary

In this chapter, we propose an unsupervised framework based on self-supervised learning with multi-target contrastive coding to learn the representations for non-native acoustic modeling of a downstream mispronunciation verification task. In our framework, the model is designed to learn the representations of the discrepancy with respect to phonetic structures in and across different languages, speakers by making predictions that are contrastive to different targets. Besides, an additional term is used to reconstruct the original speech from the shared components as a regularization. Through the experiment on Japanese part of BLCU inter-Chinese speech corpus, results show that our proposed approaches are effective to improve the performance for non-native acoustic modeling of phone recognition and mispronunciation verification.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

At the core of computer-aided language learning (CALL) systems lies the computer-aided pronunciation training (CAPT) system, designed to support language learners in improving their pronunciation, much like a language teacher would. CAPT systems play a pivotal role in enhancing learners' oral proficiency, enabling more effective and engaging language learning. However, one of the fundamental challenges in developing high-performance CAPT systems is addressing the issue of data sparsity, which significantly impacts the accuracy and reliability of such systems. In this dissertation, we have focused on advancing acoustic modeling for non-native mispronunciation verification using self-supervised learning techniques. By leveraging the power of self-supervised learning, which reduces reliance on large annotated datasets, we aimed to overcome the limitations posed by data scarcity. A series of self-supervised frameworks were proposed, tailored specifically for the task of non-native mispronunciation verification. The effectiveness of these approaches has been demonstrated through substantial improvements in experimental results, underscoring their potential to enhance the performance of CAPT systems. This work contributes to the broader field of CALL by addressing a critical gap in the development of robust and scalable pronunciation training systems. It lays the groundwork for future research in leveraging self-supervised learning for other language learning applications, ultimately paving the way for more accessible and effective language education tools.

Chapter 2 introduces a self-supervised pre-training approach for non-native mispronunciation verification. This approach aims to address the data sparsity challenge in non-native mispronunciation verification by leveraging knowledge learned from large-scale speech data in two native languages within an unsupervised framework. Specifically, unlabeled raw speech from the target language is used to train the model, enabling it

to capture phonetic properties by predicting observations within the speech data. The model is then further refined through language adversarial training using the learner's native language. This process helps the model identify patterns between the two languages, thereby enhancing its ability to generalize. We explore two adversarial training approaches: 1) an explicit auxiliary task that introduces a language discriminator to classify the language of the input sample, and 2) an implicit scheme where samples from the learner's native language are used as negative examples when calculating the prediction loss for the target language. The pre-trained model is subsequently integrated into the downstream mispronunciation verification task. We evaluate the performance of the proposed methods using the Japanese subset of the BLCU inter-Chinese speech corpus, which contains speech data from Japanese learners of Mandarin Chinese as a second language. Experimental results demonstrate that: 1) the knowledge acquired from native speech data using the proposed unsupervised framework significantly benefits both non-native phone-level speech recognition and mispronunciation verification tasks, and 2) the proposed language adversarial representation learning approach outperforms traditional methods that use non-native data alone.

Chapter 3 extends the previously proposed self-supervised pre-training framework by introducing a formant-augmented language adversarial representation learning approach for non-native mispronunciation verification. A key focus is the first convolutional layer of waveform-based convolutional encoders, which must handle high-dimensional inputs and is particularly susceptible to challenges such as the vanishing gradient problem in deep architectures. Filters learned in this layer often exhibit noisy and inconsistent multi-band shapes, a limitation that is exacerbated by insufficient training data. While these filters may function effectively within the network, they lack intuitive interpretability and may not optimally represent speech signals. To address these challenges, we propose replacing the first convolutional layer with a sinc filter to extract formant-like features. Formants, those are related to the placements and manners of articulation, are directly relevant to certain types of mispronunciations. By incorporating this information, the system can not only enhance pronunciation error detection but also provide detailed, instructive feedback to guide learners in correcting their errors. Experimental results demonstrate that integrating formant-like features via the sinc filter improves precision in both non-native phone-level speech recognition and mispronunciation veri-

fication tasks. This approach not only enhances the interpretability of the input layer but also refines the system's ability to effectively detect and address pronunciation errors. Chapter 4 investigates methods to enrich the representations learned through self-supervised training for non-native acoustic modeling. Two primary approaches are proposed: 1) multi-target contrastive coding, which enables the model to make predictions about observations that contrast with multiple targets simultaneously. This approach allows the model to learn representations that capture discrepancies in phonetic structures both within and across languages, as well as variations among speakers. 2) reconstruction regularization, which introduces an additional reconstruction term to guide the model in reconstructing the original speech from shared components. This serves as a regularization mechanism, encouraging the intermediate representations to become meaningful abstractions of the input speech. These schemes are designed to produce representations that are both transferable and semantically rich, enabling effective non-native acoustic modeling for phone recognition and mispronunciation verification without the need for human supervision. Experimental results on the Japanese subset of the BLCU inter-Chinese speech corpus show that the proposed approaches significantly improve performance in non-native acoustic modeling for both tasks.

5.2 Future Works

In this dissertation, we have proposed a series of self-supervised learning approaches for non-native mispronunciation verification and the experimental results present that the proposed approaches are effective to improve the performance of non-native mispronunciation verification and provide more informative feedback to guide the language learners to improve their pronunciation in the process of language learning. In order to apply the proposed approaches into real-world applications, more works are still needed in the future works.

- **More accurate non-native acoustic modeling for mispronunciation verification.** In this dissertation, we mainly focus to address the data sparsity problem in the establishment for mispronunciation verification system. However, there is still a significant gap between the mispronunciation verification system and the real language teachers. The performance need to be further improved to accurately

capture more mispronunciations from language learners to better guide them in their language learning.

- **Multiple native language backgrounds.** In this dissertation, we mainly focus on the establishment of mispronunciation verification system for the language learners from Japan that learns Chinese as their second language. In real world, there are a large amount of language learners from different countries with different language background. Many factors may exist to pose challenges in building mispronunciation verification system with our proposed framework. For example, the difference may be very significant between the learners' native and target languages from the aspect of articulation placements and manners, the number of samples in native or target languages may be extremely imbalanced. How to address these problems is an interesting research direction.

- **Black Box problem based on the neural network approach.** Although the neural network based problem is able to be effective enough to improve the performance of all kinds of detection systems, but if the neural network model can be further analyzed by analyzing the features learned at each. Although the neural network-based problem can effectively improve the performance of various detection systems, it would be better if the features learned in each layer of the neural network model can be further analyzed and combined with the knowledge of phonetics to know what kind of features are learned in each layer. However, if the features learned in each layer of the neural network model can be further analyzed and combined with the knowledge of phonetics to know what kind of features have been learned in each layer, it will be more effective in the application of neural networks in the computer-aided pronunciation training system, and will be more effective in helping language learners to improve the problems they encounter in the learning process.

- **Noisy scenario.** In most of the current researches, the data used consists of relatively "clean" recorded speech. However, in real-world scenarios, noise or far-field audio generally exist in speech data. It places high demands on the generalization ability of the model. Currently, most approaches involve applying various augmentation techniques to the training dataset, enabling the model to encounter

data with a certain degree of variability. However, this method incurs significant costs and offers only limited performance improvements. Therefore, it is crucial to explore ways to enhance the generalization ability of models.

Bibliography

Abdou, S. M., Hamid, S. E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., & Nazih, W. (2006). Computer aided pronunciation learning system using speech recognition techniques. *Interspeech 2006*, paper 1888–Tue1WeS.9. https://doi.org/10.21437/Interspeech.2006-287

Bouselmi, G., Fohr, D., Illina, I., & Haton, J.-P. (2006). Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints. *Ninth International Conference on Spoken Language Processing*.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108–132.

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 1–5.

Cao, W., & Zhang, J. (2009). The establishment of a capl inter-chinese corpus and its labeling. *Proceedings Of NCMMSC (in Chinese)*.

Cao, W., Wang, D., Zhang, J., & Xiong, Z. (2010). Developing a chinese l2 speech database of japanese learners with narrow-phonetic labels for computer assisted pronunciation training. *Eleventh Annual Conference of the International Speech Communication Association*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Duan, R., Kawahara, T., Dantsuji, M., & Nanjo, H. (2019). Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 391–401.

Duan, R., Zhang, J., Cao, W., & Xie, Y. (2014). A preliminary study on asr-based detection of chinese mispronunciation by japanese learners. *Fifteenth Annual Conference of the International Speech Communication Association*.

Gao, Y., Xie, Y., Cao, W., & Zhang, J. (2015). A study on robust detection of pronunciation erroneous tendency based on deep neural network. *Sixteenth Annual Conference of the International Speech Communication Association*.

Gao, Y., Xie, Y., Lin, J., & Zhang, J. (2016). Dnn based detection of pronunciation erroneous tendency in data sparse condition. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–5.

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

He, H., & Shen, X. (2007). A ranked subspace learning method for gene expression data classification. *IC-AI*, 358–364.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hu, W., Qian, Y., & Soong, F. K. (2013a). A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call). *Interspeech 2013*, 1886–1890. https://doi.org/10.21437/Interspeech.2013-458

Hu, W., Qian, Y., & Soong, F. K. (2013b). A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call). *INTERSPEECH*.

Hu, W., Qian, Y., & Soong, F. K. (2014a). A dnn-based acoustic modeling of tonal language and its application to mandarin pronunciation training. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3206–3210. https://doi.org/10.1109/ICASSP.2014.6854192

Hu, W., Qian, Y., & Soong, F. K. (2014b). A new neural network based logistic regression classifier for improving mispronunciation detection of l2 language learners. *The 9th International Symposium on Chinese Spoken Language Processing*, 245–249. https://doi.org/10.1109/ISCSLP.2014.6936712

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

Hyvarinen, A., & Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 3765–3773.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., Siebert, C., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

Jo, C.-H., Kawahara, T., Doshita, S., & Dantsuji, M. (1998). Automatic pronunciation error detection and guidance for foreign language learning. *Fifth International Conference on Spoken Language Processing*.

Joshi, S., Deo, N., & Rao, P. (2015). Vowel mispronunciation detection using dnn acoustic models with cross-lingual training. *Sixteenth Annual Conference of the International Speech Communication Association*.

Kanters, S., Cucchiarini, C., & Strik, H. (2009). The goodness of pronunciation algorithm: A detailed performance study. *Speech and Language Technology in Education (SLaTE 2009)*, 49–52. https://doi.org/10.21437/SLaTE.2009-13

Kawakami, K., Wang, L., Dyer, C., Blunsom, P., & van den Oord, A. (2020, November). Learning robust and multilingual speech representations. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1182–1192). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.106

Kim, Y., Franco, H., & Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 645–648. https://doi.org/10.21437/Eurospeech.1997-230

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koreman, J., Wik, P., Husby, O., & Albertsen, E. (2013). Universal contrastive analysis as a learning principle in capt. *Speech and Language Technology in Education*.

lang Wang, Li, C., & Meng, M. (2009). Automatic detection of phoneme-level mispronunciations. *Advanced Technology Research Bulletin*, 6–10.

Lee, A., & Glass, J. (2012). A comparison-based approach to mispronunciation detection. *2012 IEEE Spoken Language Technology Workshop (SLT)*, 382–387. https://doi.org/10.1109/SLT.2012.6424254

Lee, A., & Glass, J. (2015). Mispronunciation detection without nonnative training data. *Sixteenth Annual Conference of the International Speech Communication Association*.

Lee, A., Zhang, Y., & Glass, J. (2013). Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8227–8231. https://doi.org/10.1109/ICASSP.2013.6639269

Lin, J., Gao, Y., Zhang, W., Wei, L., Xie, Y., & Zhang, J. (2020). Improving pronunciation erroneous tendency detection with multi-model soft targets. *Journal of Signal Processing Systems*, 1–11.

Liu, Q. (2010). Research on key technology of computer-assisted mandarin pronunciation evaluation. *University of Science and Technology of China, Docotoral Thesis*, 22–28.

Maekawa, K., Kikuchi, H., & Tsukahara, W. (2004). Corpus of spontaneous japanese: Design, annotation and xml representation. *Reproduction*, *16*(16), 5–5.

Mirco, R., & Yoshua, B. (2019). Learning speaker representations with mutual information. *Proc. INTERSPEECH*, 1153–1157.

Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A., & Makino, S. (2009). A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems. *Speech Commun.*, *51*(10), 875–882. https://doi.org/10.1016/j.specom.2009.05.005

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., & Bengio, Y. (2019). Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. *Proc. INTERSPEECH*, 161–165.

Qian, X., Meng, H., & Soong, F. K. (2011). On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (capt). *Interspeech 2011*, 865–868. https://doi.org/10.21437/Interspeech.2011-330

Qian, X., Meng, H., & Soong, F. K. (2012). The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training. *Interspeech 2012*, 775–778. https://doi.org/10.21437/Interspeech.2012-238

Qian, X., Soong, F. K., & Meng, H. (2010). Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (capt). *Interspeech 2010*, 757–760. https://doi.org/10.21437/Interspeech.2010-278

Ravanelli, M., & Bengio, Y. (2018a). Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*.

Ravanelli, M., & Bengio, Y. (2018b). Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, 1021–1028.

Rivière, M., Joulin, A., Mazaré, P.-E., & Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. *arXiv preprint arXiv:2002.02848*.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). Wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Tieleman, T., & Hinton, G. (2012). Neural networks for machine learning. *Coursera (Lecture 65-RMSprop)*.

Uebler, U., & Boros, M. (1999). Recognition of non-native german speech with multilingual recognizers. *Sixth European Conference on Speech Communication and Technology*.

Viikki, O., & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*(1-3), 133–147.

Wang, Y.-B., & Lee, L.-S. (2012). Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. *2012*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5049–5052.

Wang, Z., Schultz, T., & Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, *1*, I–I.

Wei, S., Hu, G., Hu, Y., & Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, *51*(10), 896–905.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, *14*(4), 715–770.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2-3), 95–108.

Witt, S. M. (1999). Use of speech recognition in computer-assisted language learning.

Wu, Z., & Lin, M. (1989). *Summary of Experimental Phonetics*.

Yang, L., Fu, K., Zhang, J., & Shinozaki, T. (2020). Pronunciation erroneous tendency detection with language adversarial represent learning.

Yang, L., Fu, K., Zhang, J., & Shinozaki, T. (2021). Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning. *Neural Networks*, *142*, 597–607. https://doi.org/https://doi.org/10.1016/j.neunet.2021.07.017

Yang, L., Fu, K., Zhang, J., & Shinozaki, 0. (2021). Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning. *Neural Networks*, *142*, 597–607.

Yang, L., Xie, Y., Gao, Y., & Zhang, J. (2017). Improving pronunciation erroneous tendency detection with convolutional long short-term memory. *2017 International Conference on Asian Language Processing (IALP)*, 52–56.

Yang, L., Zhang, J., & Shinozaki, T. (2022). Self-supervised learning with multi-target contrastive coding for non-native acoustic modeling of mispronunciation verification. *Interspeech 2022*, 4312–4316. https://doi.org/10.21437/Interspeech.2022-207

Yuan, H., Zhao, J., & Liu, J. (2012). Improve mispronunciation detection with tandem feature. *2012 8th International Symposium on Chinese Spoken Language Processing*, 184–187. https://doi.org/10.1109/ISCSLP.2012.6423538

Zhang, F., Huang, C., Soong, F. K., Chu, M., & Wang, R. (2008). Automatic mispronunciation detection for mandarin. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5077–5080. https://doi.org/10.1109/ICASSP.2008.4518800

Zheng, J., Huang, C., Chu, M., Soong, F. K., & Ye, W.-p. (2007). Generalized segment posterior probability for automatic mandarin pronunciation evaluation. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, *4*, IV–201.