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
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Information and Communication Engineering 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineering)
学生氏名： Student's Name	Longfei Yang	審査員主査： Chief Examiner	Takahiro Shinozaki

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

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: 1) Multi-target contrastive coding: This approach contrasts phonetic discrepancies both within and across languages and speakers, enabling the model to learn nuanced phonetic representations. 2) 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.