Comparative Analysis of Acoustic Perception Models in Simulation of Teacher-Learner Interaction in L2 Pronunciation Learning

Anonymous ACL submission

Abstract

This study presents a comparative analysis of acoustic perception models in simulating teacher-learner interaction for second language (L2) English pronunciation learning, focusing on Chinese native speakers. Three acoustic perception models are evaluated: an English model (M1) based on the XLS-R framework and fine-tuned on the TIMIT corpus, a non-native model (M2) also based on XLS-R but fine-tuned on the L2-ARCTIC corpus, and a Chinese model (M3) using a sequence-to-sequence architecture with connectionist temporal classification (CTC) fine-tuned on the AISHELL-1 corpus. A corpus of seven pseudo-words designed to challenge Chinese learners of English is used to assess the models' performance in capturing the acoustic perception of L2 learners. The Levenshtein distance between recognised sequences and reference sequences for Chinese and English speakers is employed as an evaluation metric, along with the ratio of these distances. Results show that the non-native model (M2) outperforms the English (M1) and Chinese (M3) models in minimising the Levenshtein distance for Chinese speakers and achieves the lowest ratio, indicating its effectiveness in modelling the acoustic perception of L2 learners. These findings suggest that incorporating non-native speech data in acoustic perception models can improve the simulation of teacher-learner interaction in L2 pronunciation learning.

1 Introduction

Acoustic perception plays a crucial role in second language (L2) pronunciation learning, as it directly influences learners' ability to accurately perceive and produce sounds in the target language (Mitterer and Ernestus, 2008). Computational modelling of L2 acoustic perception offers valuable insights into the underlying processes and challenges faced by learners, enabling the development of more effective language learning technologies and pedagogical approaches. While several approaches have been proposed for modelling L2 acoustic perception, such as using native speech models (Kanters et al., 2009; Witt and Young, 2000) or specialised L2 acoustic models (Franco et al., 2010; Li et al., 2016), these methods have limitations in capturing the specific challenges faced by L2 learners from different first language (L1) backgrounds. Native speech models may not fully account for the perceptual difficulties experienced by L2 learners, while specialised L2 models often require large amounts of L2 speech data, which may not be readily available for all language pairs or proficiency levels. This study addresses this research gap by investigating the effectiveness of non-native acoustic perception models, specifically for Chinese learners of English. This work advances the state-of-the-art in computational modelling of L2 acoustic perception by leveraging self-supervised models like XLS-R (Baevski et al., 2020) and fine-tuning them on native and non-native speech data. The novel approach of focusing on a specific L1 background (Chinese) allows for a more targeted evaluation of the models' performance and provides insights into the perceptual patterns of this learner population. The findings of this study demonstrate that non-native acoustic perception models outperform native and L1-specific models in capturing the perceptual patterns of Chinese learners of English. Specifically, the results show that a model fine-tuned on non-native speech data (L2-ARCTIC corpus) achieves the lowest Levenshtein distance and ratio when compared to native English and Chinese models, indicating its effectiveness in modelling the acoustic perception of Chinese L2 learners. These results have significant implications for the development of personalised language learning technologies and inform pedagogical approaches for L2 pronunciation training. The remainder of this paper is organised as follows: Section 2 provides an overview of L2 acoustic perception and its challenges, followed by a discussion of computational acoustic perception models in Section 3. Section 4 describes the simulation of teacher-learner interaction, and Section 5 details the implementation of the acoustic perception models used in this study. The corpus and evaluation methodology are presented in Section 6, followed by the results and discussion in Section 7. Sections 8 and 9 compare our approach with other methods and discuss the limitations of the study. Finally, Section 10 concludes the paper and outlines future research directions.

2 L2 Learners Acoustic Perception

Understanding acoustic perception is crucial in L2 learning due to its direct influence on production (Mitterer and Ernestus, 2008). L2 learners must learn to perceive and produce sounds that may not exist or may be articulated differently in their first language (L1). Many challenges faced by L2 learners in accurately perceiving and producing the sounds of the target language arise from differences in phonetic systems and phonological rules between their L1 and the L2. For example, English speakers learning Chinese may struggle to distinguish between the contrasting phonemic tones (Hao, 2012), while Chinese speakers learning English may have difficulty differentiating between similar English consonant sounds like /r/ and /l/ (Radant et al., 2009).

Another aspect of acoustic perception is the ability to perceive and produce correct intonation patterns and stress. Intonation plays a crucial role in conveying meaning and pragmatic nuances in speech, and L2 learners need to develop sensitivity to the intonational contours of the target language. Stress patterns also vary across languages, and learners must learn to identify and reproduce the appropriate stress patterns to convey meaning accurately (Liu and Reed, 2021; Braun et al., 2014; Altmann, 2006).

Furthermore, L2 learners may face challenges related to the rhythm and timing of speech. Languages vary in their rhythmic patterns, with some languages exhibiting syllable-timed rhythm (e.g., French, Spanish) and others stress-timed rhythm (e.g., English) (Barry, 2007; Ordin and Polyanskaya, 2015). Various factors influence the development of acoustic perception in L2 learners, including age of learning, exposure to the target language, individual differences in auditory processing abilities, and instructional methods (Saito et al., 2024). Effective learning strategies for improving acoustic perception in L2 learners include explicit phonetic instruction, focused listening practice, auditory discrimination tasks, and feedback on pronunciation accuracy (Kissling, 2015). Experimental research methods using psychophysical techniques to measure L2 acoustic perception involve conducting controlled experiments with human participants to investigate how they perceive and process acoustic features of a second language (Sakai and Moorman, 2018). These methods include discrimination tasks, wherein participants are presented with pairs of stimuli (e.g., pairs of phonemes or words) that differ along some acoustic dimension. Subsequently, participants are asked to indicate whether the stimuli in each pair are the same or different (Aliaga-García and Mora, 2009; Zhen and Pratt, 2023). Additionally, ABX tasks involve presenting participants with three stimuli (A, B, and X), where A and B are similar stimuli, and X is either identical to A or B. Participants are then asked to indicate whether X matches A or B. This task aids in assessing discrimination abilities while controlling for perceptual biases (Greenaway, 2017; Melnik-Leroy et al., 2022). Despite the insights provided by experimental research using psychophysical techniques, these methods have limitations in fully capturing the complexity of L2 acoustic perception. Various factors influence the development of acoustic perception in L2 learners, posing challenges for comprehensive consideration in experimental settings. Furthermore, experimental methods employing psychophysical techniques may require participants to make fine-grained judgments about subtle acoustic differences, demanding significant cognitive effort and potentially failing to fully capture participants' naturalistic perception of L2 speech (Leow, 2015). Consequently, researchers have increasingly turned to computational acoustic perception models to address these limitations and provide deeper analysis into L2 speech perception.

3 Computational Acoustic Perception Models

Computational acoustic perception models aim to replicate how humans perceive and process sound (Jepsen et al., 2008; Kröger et al., 2009). These models integrate principles from signal processing and neuroscience to understand and interpret acoustic signals. The computational auditory signal-processing and perception model (CASP) (Jepsen et al., 2008) is an adaptation of an earlier model developed by Dau et al. in 1997 (Dau et al., 1997). The CASP model includes an outer and middle-ear transformation, along with a nonlinear cochlear filtering stage known as the dual resonance nonlinear (DRNL) filterbank, which replaces the linear gammatone filterbank used in the original model. The DRNL filterbank better captures the nonlinear processing that occurs in the human cochlea, allowing for more accurate modelling of auditory perception. Other computational models of auditory perception, such as those proposed by Meddis and O'Mard (1997), Zilany and Bruce (2006), and Mao and Carney (2015), aim to construct comprehensive models that capture essential acoustic features. These models can be utilised to explore various aspects of human auditory perception, including pitch perception, temporal processing, and the perception of complex sounds. In addition to models specifically designed for auditory perception, many speech models, such as automatic speech recognition (ASR) and speech synthesis, employ acoustic feature models that focus on extracting relevant features from acoustic signals crucial for perception.

Recent advancements in self-supervised speech representation learning (Close et al., 2023; Mohamed et al., 2022) have enabled these models to autonomously learn to discern acoustic features and categorise or predict various perceptual attributes without the need for explicit labelling. By training on large, task-specific corpora, these self-supervised models can capture rich representations of speech that are useful for a wide range of downstream tasks, including acoustic perception modelling. Furthermore, other work, such as that by Islam et al. (2023), utilises automatic phoneme recognition as an acoustic perception model. This approach leverages the ability of phoneme recognition systems to identify and classify individual speech sounds, providing a framework for modelling the perception of phonetic units in speech. The details of this work and its implications for L2 pronunciation learning will be elaborated on in the following section.

Overall, computational acoustic perception models offer a powerful tool for understanding and simulating human auditory perception. By combining insights from signal processing, neuroscience, and machine learning, these models can provide valuable insights into the mechanisms underlying acoustic perception and inform the development of more effective strategies for L2 pronunciation learning.

4 Simulation of Teacher-Learner Interaction

The system design depicted in Figure 1 presents the general framework of teacher and learner interaction in English pronunciation learning, as introduced in (Islam et al., 2023), inspired by studies on speech learning models (Bohn and Munro, 2007; Flege and Bohn, 2021). The system is divided into two parts: the teacher model and the learner model, each consisting of multiple submodels. This work will focus on different implementations of the acoustic perception model in the learner model.

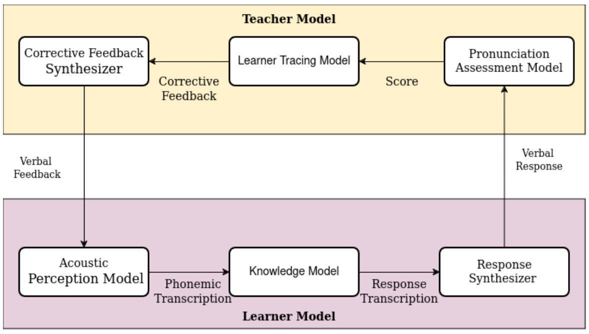

Figure 1: The model for teacher-learner interaction in English pronunciation learning by Islam et al. (2023).

4.1 Teacher Model

Inspired by research in teacher modelling (Shadiev and Yang, 2020; Slavuj et al., 2015), the teacher model in Figure 1 simulates an English native speaker and employs repetition as a teaching strategy. The teacher model is composed of a pronunciation assessment model, which receives the verbal response from the learner model, assesses the pronunciation, and sends the score to the feedback generator model. The pronunciation assessment model is implemented as Goodness Of Pronunciation (GOP), which was initially introduced by Kim et al. (1997) and improved by Sudhakara et al. (2019), using the Kaldi tool (Povey et al., 2011) trained on the WSJCAM0 British English corpus (Robinson et al., 1995). The learner tracing model aims to understand how the learner model engages through the learning process. It is a rule-based model that takes the score as input, updates the learner state, and generates feedback based on the learner state. The corrective feedback synthesizer model generates verbal feedback using an end-to-end text-to-speech synthesis model, trained using Fastspeech2 (Ren et al., 2020) with a publicly available English speech corpus LJ that comprises 13,100 short audio clips by a single speaker (Ito, 2017).

4.2 Learner Model

The learner model in Figure 1 simulates a Chinese-speaking learner in the early stages of English learning. As the learner engages with the teacher model, verbal feedback is perceived and processed into phonemic transcriptions by an acoustic perception model, which will be discussed in more detail in Section 5. The second model is the knowledge model, which is a rule-based model informed by the perceived phonemic transcriptions and updated by a learner knowledge model to generate the response. The response synthesizer employs Fastspeech2 (Ren et al., 2020) trained on the AISHELL-3 corpus (Shi et al., 2020).

5 Acoustic Perception Model Implementation

The main goal of the acoustic perception model in Figure 1 is to emulate the perception of non-native speakers. This is implemented in the form of phone recognises trained on different languages, which are bound to make errors when trying to recognise the verbal feedback of teachers, in this case, the feedback from the teacher model in the English language. These models are built for capturing a broad spectrum of acoustic features, encompassing phonetic variations, prosodic patterns, intonations, spectral characteristics, temporal dynamics, and the articulatory differences that exist between languages, even on identical phonemic representations.

Word-level automatic speech recognition (ASR) focuses on recognising entire words, whereas phoneme-level recogniser focus on recognising individual phonetic units. Phoneme-level recognises are generally more flexible when dealing with speech in different languages or with unfamiliar words, as they operate at a more fundamental level of linguistic representation. The performance of these models is evaluated in terms of phone error rate (PER), which is the percentage of incorrectly recognised phone sequences in relation to the reference recognised phone sequences. The use of phoneme-level recogniser as acoustic perception models in the simulation of teacher-learner interaction offers several advantages. By modelling perception at the phonetic level, these recogniser can capture the fine-grained acoustic differences between the learner's native language and the target language. This allows for a more accurate representation of the challenges faced by non-native speakers in perceiving and processing the sounds of the target language.

Furthermore, by training these recogniser on different languages, the acoustic perception model can simulate the influence of the learner's native language on their perception of the target language. This is particularly relevant in the context of Chinese-speaking learners acquiring English pronunciation, as the phonetic inventories and phonological rules of these two languages differ significantly. The incorporation of phoneme-level recogniser as acoustic perception models in the learner model enables a more realistic simulation of the perceptual processes involved in L2 pronunciation learning. By capturing the errors and variations in phonetic perception, this approach can provide valuable insights into the challenges faced by non-native speakers and inform the development of more effective teaching strategies and feedback mechanisms in the teacher model.

5.1 English Acoustic Perception Model

The English acoustic perception model $(M1)$ is based on a large-scale pretrained foundation model called XLS-R, which aims at cross-lingual speech representation learning and builds upon the Wav2Vec 2.0 framework (Baevski et al., 2020). XLS-R has been trained on a diverse corpus encompassing 53 languages, totaling 56k hours of speech data sourced from CommonVoice (Ardila et al., 2019), BABEL (Roach et al., 1996), and Multilingual LibriSpeech (Pratap et al., 2020). The English acoustic perception model is fine-tuned on the TIMIT corpus (Garofolo et al., 1993), which is widely used for speech recognition and linguistic research purposes. The TIMIT corpus contains phoneme-level transcriptions and recordings of 630 speakers from various regions of the United States, representing a diversity of demographics such as gender, race, and age. The model obtained a PER of 7.996%.

5.2 Non-Native Acoustic Perception Model

The architecture of the non-native acoustic perception model $(M2)$ is also based on the large-scale pretrained foundation model XLS-R. However, the model is now fine-tuned on the L2-ARCTIC corpus, which is specifically tailored for non-native English speech research (Zhao et al., 2018). The L2-ARCTIC corpus contains phoneme-level transcriptions and comprises recordings from 24 non-native speakers of English, originating from diverse language backgrounds including Hindi, Korean, Chinese, Spanish, Arabic, and Vietnamese. Within each language group, recordings are available from two male and two female speakers, ensuring a balanced representation across genders and languages. The model achieved a PER of 12.8%.

5.3 Chinese Acoustic Perception Model

The Chinese acoustic perception model $(M3)$ is a sequence-to-sequence model with connectionist temporal classification (CTC) based on a universal phone recognizer that aims to deploy recognition with a multilingual (universal) allophone system (Li et al., 2020). It was trained on data from eleven languages, including English, Japanese, Chinese, and Tagalog. The model architecture consists of a bidirectional Long Short-Term Memory (LSTM) encoder (Malhotra et al., 2015). The acoustic model is fine-tuned on AISHELL-1, an open-source Chinese speech corpus (Bu et al., 2017) containing 400 speakers and over 170 hours of Chinese speech data. Since AISHELL-1 does not contain phoneme-level transcriptions, Kaldi (Povey et al., 2011) speech recognition tools are used to extract the time alignment to obtain phoneme transcriptions. The corpus with the phoneme transcription is then used to fine-tune the model (Li et al., 2020). The model obtained a PER of 22.3%.

These three acoustic perception models, representing native English, non-native English, and Chinese perception, offer a diverse set of tools for simulating and understanding the challenges faced by learners in perceiving and processing the sounds of the target language. By incorporating these models into the learner model of the teacher-learner interaction framework, researchers can gain valuable insights into the perceptual processes involved in L2 pronunciation learning and develop more effective teaching strategies and feedback mechanisms.

6 Corpus Description

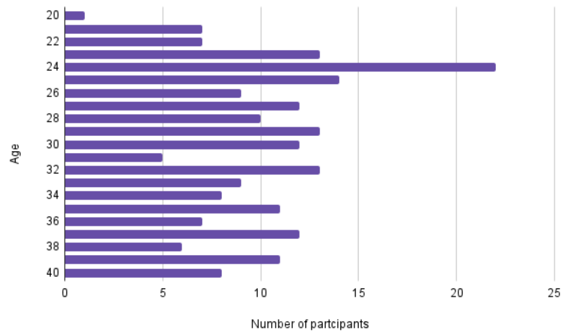

Figure 2: The distribution of participant's age. The age range from 20 to 40.

6.1 Pseudo-Word Design and Validation

In examining L2 pronunciation perception and production, a particular group of seven pseudo-words was selected to ensure that they were not influenced by written forms or any prior knowledge of pronunciation. Each word consists of 6-7 phonemes. The words were created to include phonemes that are known to be difficult for Chinese learners, such as /l/, /r/, /ʃ/, /g/, /v/, and /ð/ (Zhang and Xiao, 2014; Richards, 2011). Two experienced English pronunciation instructors were consulted to validate the suitability of the pseudo-words for the study. A pilot test with five Chinese learners of English was conducted to ensure that the words were challenging but not impossible to pronounce. The experimental word list is presented in Table 1 along with with IPA transcription. [1].

6.2 Data Collection

The study involved 240 participants, including 120 Chinese native speakers (ChS) and 120 English native speakers (EnS). Among them, 150 participants were aged between 20 and 30, while 90 participants were aged between 31 and 40. The data collection process was facilitated through a dedicated website designed specifically for this purpose.

For each of the seven pseudo-words, participants listened to the corresponding audio file and selected the correct answer from three audio options.

[1] IPA (International Phonetic Alphabet) is a standardised set of symbols used to represent the sounds of spoken language, providing a clear and accurate way to transcribe the pronunciation of words across different languages.

Table 1: Experimental word list with word ID, pseudo-words, IPA transcription, and comments on the challenging phonemes marked in the last column.

Word ID	**Pseudo-words**	**IPA**	**Comment**
w_1	RALISAR	/rælɪsɑr/	The inability to distinguish /l/ from /r/ and /æ/ from /ɑr/.
w_2	SHEEBINGS	/ʃiːbɪŋ/	The /ʃ/ and the /g/ sound.
w_3	BADUNLOT	/bædʌnlɒt/	The final /t/ often becomes a glottal stop [ʔ], so the word may be recognised when read but difficult to identify in spoken language.
w_4	MASIGAN	/mæsɪgæn/	The /æ/ sound and the /g/ sound.
w_5	NAVIKLY	/nævɪkliː/	The /v/ sound and often use /w/ instead.
w_6	TAGAMAUGH	/tægæmɑːf/	Words ending in "ugh" are sometimes a diphthong (e.g., though /ðəʊ/) but could be the sound /f/.
w_7	HICKOMAY	/hɪkʌmeɪ/	The diphthong /eɪ/. The weak vowel /ʌ/ is in the middle of the word.

6.2.1 Evaluation Metrics

The Levenshtein distance (Levenshtein et al., 1966) was used to measure the similarity between the recognised phoneme sequences and the reference sequences for both ChS and EnS speakers. This distance calculates the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. A lower Levenshtein distance indicates a higher similarity between the sequences.

To compare the models' performance in relation to EnS, the ratio of the average Levenshtein distance for ChS to the average Levenshtein distance for EnS was calculated. A lower ratio suggests that the model better captures the acoustic perception of ChS relative to EnS.

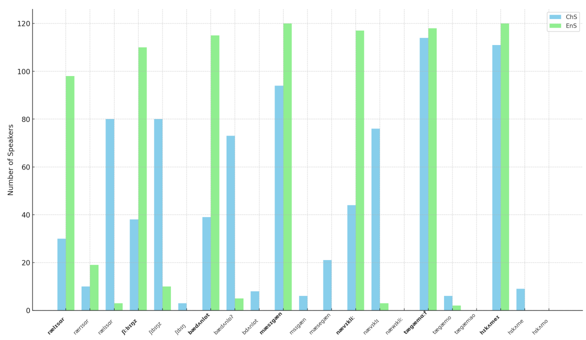

Figure 3: Occurrence counts of different IPA transcripts for pseudo-words. Each pair of bars represents the number of speakers, differentiated by color: skyblue for Chinese speakers (ChS) and lightgreen for English speakers (EnS). The IPA transcript in bold denotes the correct answer.

Let M_1, M_2, and M_3 be the three acoustic perception models to be evaluated. For each pseudo-word w_i, $i \in 1, 2, ..., 7$, the most frequently selected answer among the three options will serve as the reference sequence, denoted as r_i. The Levenshtein distance (Levenshtein et al., 1966) between the recognised sequence by model M_j, $j \in 1, 2, 3$, for pseudo-word w_i and the reference sequence r_i of EnS is calculated as:

$$LD_{ChS}(M_j, w_i) = \frac{1}{n} \sum_{k=1}^{n} d(s_{jk}, r_i) \quad (1)$$

where n is the number of speakers in ChS, s_{jk} is the recognised sequence by model M_j for speaker k in ChS, and $d(\cdot, \cdot)$ is the Levenshtein distance function. Similarly, the averaged phoneme distance between the recognised sequence by model M_j for pseudo-word w_i and the reference sequence of EnS is calculated as:

$$LD_{EnS}(M_j, w_i) = \frac{1}{m} \sum_{k=1}^{m} d(s_{jk}, r_i) \quad (2)$$

where m is the number of speakers in EnS. The ratio between the two distances for model M_j and pseudo-word w_i is then calculated as:

$$R(M_j, w_i) = \frac{LD_{ChS}(M_j, w_i)}{LD_{EnS}(M_j, w_i)} \quad (3)$$

The average ratio across all pseudo-words for model M_j is:

$$\bar{R}(M_j) = \frac{1}{7} \sum_{i=1}^{7} R(M_j, w_i) \quad (4)$$

The acoustic perception model with the lowest average ratio $\bar{R}(M_j)$ will be considered the most suitable for integration with the simulation model, as it maximises the similarity with ChS over EnS.

6.2.2 Results

The results demonstrate that the M_2 outperforms M_1 and M_3 models in capturing the perceptual patterns of Chinese learners of English. Across all pseudo-words, M_2 achieves the lowest average Levenshtein distance for Chinese speakers (2.87) and the lowest ratio (0.69) between the distances for Chinese and English speakers (Table 3).

A closer examination of the results reveals that M_2 is particularly effective in modelling the perception of challenging phonemes for Chinese learners. For example, in pseudo-word w_1 (/rælɪsɑr/), M2 achieves a Levenshtein distance of 0.22 for Chinese speakers, compared to 0.42 for M_1 and 0.48 for M_3 (Tables 2, 3, 4). This suggests that M_2 better captures the difficulty Chinese learners face in distinguishing between /l/ and /r/ sounds. Similarly, for pseudo-word w_5 (/nævɪkli/), M_2 achieves a distance of 0.33 for Chinese speakers, while M_1 and M_3 have distances of 0.42 and 0.51, respectively. This indicates that M_2 is more sensitive to the challenges Chinese learners encounter with the /v/ sound, which is often substituted with /w/.

Across all pseudo-words, M_2 achieves the lowest average $LD_{ChS}(M_j, w_i)$ of 2.87, compared to 3.56 for M_1 and 4.5 for M_3. Furthermore, the average ratio $\bar{R}(M_j)$ provides insight into how well each model captures the acoustic perception of Chinese speakers relative to English speakers. A lower ratio indicates better performance in modelling the acoustic perception of Chinese speakers. M_2 has the lowest ratio at 0.69, followed by M_1 at 0.73 and M_3 at 0.85. Considering both the average $LD_{ChS}(M_j, w_i)$ and the $\bar{R}(M_j)$ ratio, the non-native acoustic perception model (M_2) emerges as the best choice for simulating the acoustic perception of Chinese speakers learning English.

7 Comparison with Other Approaches

Several approaches have been proposed for modelling acoustic perception in L2 pronunciation learning, each with its own strengths and limitations. One common approach is the use of The goodness of pronunciation algorithm (GOP) trained on native speech data to evaluate L2 learners' pronunciations (Kanters et al., 2009; Witt and Young,

Table 2: Performance measures for the English acoustic perception model (M_1). The Levenshtein distances between the recognised sequences and the reference sequences for Chinese speakers (ChS) and English speakers (EnS) are denoted as $LD_{ChS}(M_1, w_i)$ and $LD_{EnS}(M_1, w_i)$, respectively. The average distances across all pseudo-words and the ratio $\bar{R}(M_1)$ are also provided.

Word	$LD_{ChS}(M_1, w_i)$	$LD_{EnS}(M_1, w_i)$
w_1	0.42	0.85
w_2	0.36	0.75
w_3	0.39	0.73
w_4	0.63	0.63
w_5	0.42	0.57
w_6	0.73	0.73
w_7	0.61	0.61
Average	3.56	4.87
$\bar{R}(M_1)$	0.73	

Table 3: Performance measures for the non-native acoustic perception model (M_2). The Levenshtein distances between the recognised sequences and the reference sequences for Chinese speakers (ChS) and English speakers (EnS) are denoted as $LD_{ChS}(M_2, w_i)$ and $LD_{EnS}(M_2, w_i)$, respectively. The average distances across all pseudo-words and the ratio $\bar{R}(M_2)$ are also provided.

Word	$LD_{ChS}(M_2, w_i)$	$LD_{EnS}(M_2, w_i)$
w_1	0.22	0.65
w_2	0.26	0.52
w_3	0.29	0.63
w_4	0.51	0.51
w_5	0.33	0.58
w_6	0.61	0.61
w_7	0.65	0.65
Average	2.87	4.15
$\bar{R}(M_2)$	0.69	

Table 4: Performance measures for the Chinese acoustic perception model (M_3). The Levenshtein distances between the recognised sequences and the reference sequences for Chinese speakers (ChS) and English speakers (EnS) are denoted as $LD_{ChS}(M_3, w_i)$ and $LD_{EnS}(M_3, w_i)$, respectively. The average distances across all pseudo-words and the ratio $\bar{R}(M_3)$ are also provided.

Word	$LD_{ChS}(M_3, w_i)$	$LD_{EnS}(M_3, w_i)$
w_1	0.48	0.80
w_2	0.53	0.75
w_3	0.57	0.69
w_4	0.77	0.77
w_5	0.51	0.61
w_6	0.78	0.78
w_7	0.86	0.86
Average	4.5	5.26
$\bar{R}(M_3)$	0.85	

2000). While this approach provides a straightforward way to assess pronunciation quality, it may not fully capture the specific challenges faced by L2 learners, as it relies on models trained on native speech patterns.

Another approach is the use of specialised acoustic models trained on L2 speech data (Franco et al., 2010; Li et al., 2016). These models are designed to capture the specific acoustic characteristics of L2 learners' speech and have been shown to improve the performance of pronunciation assessment systems. However, these models often require a large amount of L2 speech data, which may not always be available for all language pairs or proficiency levels. In contrast, our proposed approach leverages pre-trained, self-supervised models like XLS-R, which are trained on a large amount of multilingual speech data. By fine-tuning these models on smaller amounts of native and non-native speech data, we can create acoustic perception models that are better suited to capturing the perceptual challenges faced by L2 learners.

8 Conclusions

This study demonstrates the importance of considering non-native speech data when developing acoustic perception models for simulating teacher-learner interaction in L2 English pronunciation learning. By comparing the performance of native English, non-native, and Chinese acoustic perception models, it found that the non-native model M_2 fine-tuned on the L2-ARCTIC corpus outperformed the other models in capturing the perceptual patterns of Chinese learners of English. This finding highlights the effectiveness of incorporating non-native speech data in modelling L2 acoustic perception. The superior performance of the non-native acoustic perception model has significant implications for L2 pronunciation teaching and learning. By incorporating models like M2 into CAPT systems, we can develop more effective tools that provide targeted feedback to Chinese learners of English. For instance, when a learner mispronounces a word containing /l/ or /r/, the system can identify the specific error and offer personalised guidance on how to produce the correct sound. This can lead to more efficient and engaging pronunciation practice, as learners receive immediate and relevant feedback. Future research should build upon these findings by investigating the performance of non-native acoustic perception models with a more diverse range of L2 learners, expanding the corpus to include a larger variety of words and phonemes, and exploring additional evaluation metrics. Moreover, integrating these acoustic perception models into a complete simulation framework of teacher-learner interaction would provide a more comprehensive understanding of their impact on L2 pronunciation learning. In conclusion, this study underscores the potential of non-native acoustic perception models in advancing computational modelling of L2 speech perception and informing the development of effective language learning technologies. As research in this field continues to progress, the insights gained from this work can contribute to creating more adaptive and personalised tools to support L2 learners in their pronunciation learning journey.

9 Preserving Anonymity and Ethics

Participants received Participant Information Sheets and Consent Forms approved by the University Research Ethics Committee. These documents outlined project details, stressed voluntary participation, and provided withdrawal options. The university ensured secure, anonymous data storage and transportation, retaining anonymised data for at least 10 years post-study.

10 Limitations

While this study provides valuable insights into the effectiveness of different acoustic perception models for simulating teacher-learner interaction in L2 English pronunciation learning, there are several limitations to consider.

First, the study focuses on a specific group of learners, Chinese native speakers, and the findings may not generalise to learners from other language backgrounds. Future research should investigate the performance of these models with a more diverse range of L2 learners.

Second, the corpus used in this study consists of a limited number of pseudo-words, which may not fully capture the complexity of English pronunciation. Expanding the corpus to include a larger variety of words and phonemes could provide a more comprehensive evaluation of the models' performance.

Third, the study relies on the Levenshtein distance as the primary evaluation metric, which may not fully capture the nuances of acoustic perception. Incorporating additional metrics, such as phoneme confusion matrices (Leijon et al., 2015), could provide a more comprehensive assessment of the models' performance.

Finally, the study does not address the integration of these acoustic perception models into a complete simulation of teacher-learner interaction. Future work should investigate how these models can be incorporated into a larger framework that includes other components, such as feedback generation and learner modelling, to provide a more comprehensive simulation of L2 pronunciation learning.

11 Acknowledgements

References

Cristina Aliaga-García and Joan C Mora. 2009. Assessing the effects of phonetic training on l2 sound perception and production. *Recent research in second language phonetics/phonology: Perception and production*, 231.

Heidi Altmann. 2006. *The perception and production of second language stress: A cross-linguistic experimental study*. University of Delaware.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

William J Barry. 2007. Rhythm as an l2 problem: How prosodic is it. *Non-native prosody: Phonetic description and teaching practice*, pages 97–120.

Ocke-Schwen Bohn and Murray J Munro. 2007. *Language experience in second language speech learning: In honor of James Emil Flege*, volume 17. John Benjamins Publishing.

Bettina Braun, Tobias Galts, and Barış Kabak. 2014. Lexical encoding of l2 tones: The role of l1 stress, pitch accent and intonation. *Second Language Research*, 30(3):323–350.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

George Close, Thomas Hain, and Stefan Goetze. 2023. The effect of spoken language on speech enhancement using self-supervised speech representation loss functions. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. 1997. Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5):2892–2905.

James Emil Flege and Ocke-Schwen Bohn. 2021. The revised speech learning model (slm-r). *Second language speech learning: Theoretical and empirical progress*, pages 3–83.

Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda. 2010. Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401–418.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Ruth Elizabeth Greenaway. 2017. Abx discrimination task. In *Discrimination Testing in Sensory Science*, pages 267–288. Elsevier.

Yen-Chen Hao. 2012. Second language acquisition of mandarin chinese tones by tonal and non-tonal language speakers. *Journal of phonetics*, 40(2):269–279.

Elaf Islam, Thomas Hain, and Protima Nomo Sudro. 2023. Simulation of teacher-learner interaction in english language pronunciation learning. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Keith Ito. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Morten L Jepsen, Stephan D Ewert, and Torsten Dau. 2008. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438.

Sandra Kanters, Catia Cucchiarini, and Helmer Strik. 2009. The goodness of pronunciation algorithm: a detailed performance study.

Yoon Kim, Horacio Franco, and Leonardo Neumeyer. 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In *Fifth European Conference on Speech Communication and Technology*.

Elizabeth M Kissling. 2015. Phonetics instruction improves learners' perception of l2 sounds. *Language teaching research*, 19(3):254–275.

Bernd J Kröger, Jim Kannampuzha, and Christiane Neuschaefer-Rube. 2009. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809.

Arne Leijon, Gustav Eje Henter, and Martin Dahlquist. 2015. Bayesian analysis of phoneme confusion matrices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):469–482.

Ronald P Leow. 2015. *Explicit learning in the L2 classroom: A student-centered approach*. Routledge.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Di Liu and Marnie Reed. 2021. Exploring the complexity of the l2 intonation system: An acoustic and eye-tracking study. *Frontiers in Communication*, 6:627316.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal, et al. 2015. Long short term memory networks for anomaly detection in time series. In *Esann*, volume 2015, page 89.

Junwen Mao and Laurel H Carney. 2015. Tone-in-noise detection using envelope cues: comparison of signal-processing-based and physiological models. *Journal of the Association for Research in Otolaryngology*, 16:121–133.

Ray Meddis and Lowel O'Mard. 1997. A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, 102(3):1811–1820.

Gerda Ana Melnik-Leroy, Rory Turnbull, and Sharon Peperkamp. 2022. On the relationship between perception and production of l2 sounds: Evidence from anglophones' processing of the french/u/–/y/contrast. *Second Language Research*, 38(3):581–605.

Holger Mitterer and Mirjam Ernestus. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109(1):168–173.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*.

Mikhail Ordin and Leona Polyanskaya. 2015. Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *The Journal of the Acoustical Society of America*, 138(2):533–544.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Hui-Ling Huang& James Radant, H James, and H Huang. 2009. Chinese phonotactic patterns and the pronunciation difficulties of mandarin-speaking efl learners. *The Asian EFL Journal Quarterly*, 11(4):115.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Monica Richards. 2011. Helping chinese learners distinguish english/l/and/n. *Pronunciation in Second Language Learning and Teaching Proceedings*, 3(1).

Peter Roach, Simon Arnfield, William Barry, Julia Baltova, Marian Boldea, Adrian Fourcin, Wiktor Gonet, Ryszard Gubrynowicz, Elisabeth Hallum, Lori Lamel, et al. 1996. Babel: An eastern european multi-language database. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1892–1893. IEEE.

Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. 1995. Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 81–84. IEEE.

Kazuya Saito, Magdalena Kachlicka, Yui Suzukida, Ingrid Mora-Plaza, Yaoyao Ruan, and Adam Tierney. 2024. Auditory processing as perceptual, cognitive, and motoric abilities underlying successful second language acquisition: Interaction model. *Journal of Experimental Psychology: Human Perception and Performance*, 50(1):119.

Mari Sakai and Colleen Moorman. 2018. Can perception training improve the production of second language phonemes? a meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, 39(1):187–224.

Rustam Shadiev and Mengke Yang. 2020. Review of studies on technology-enhanced language learning and teaching. *Sustainability*, 12(2):524.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Vanja Slavuj, Božidar Kovačić, and Igor Jugo. 2015. Intelligent tutoring systems for language learning. In *2015 38th MIPRO*, pages 814–819. IEEE.

Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2019. An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities. In *INTERSPEECH*, volume 2, pages 954–958.

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

Yanyan Zhang and Jing Xiao. 2014. An analysis of chinese students' perception and production of paired english fricatives: From an elf perspective. *Journal of Pan-Pacific Association of Applied Linguistics*, 18(1):171–192.

Guanlong Zhao, Evgeny Chukharev-Hudilainen, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Ricardo Gutierrez-Osuna, and John Levis. 2018. L2-arctic: A non-native english speech corpus.

Leslie Q Zhen and Sheila R Pratt. 2023. Perceptual, procedural, and task learning for an auditory temporal discrimination task. *The Journal of the Acoustical Society of America*, 153(3):1823–1835.

Muhammad SA Zilany and Ian C Bruce. 2006. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, 120(3):1446–1466.