SIMULATION OF TEACHER-LEARNER INTERACTION IN ENGLISH LANGUAGE PRONUNCIATION LEARNING

Elaf Islam, Thomas Hain, Protima Nomo Sudro

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK
{ejislam1, t.hain, p.nomo.sudro}@sheffield.ac.uk

ABSTRACT

Second language (L2) learning is a complex process that is difficult to model. This work aims to develop a computational model of the teacher–learner interaction as used for L2 learning. The teacher model simulates a native English speaker, which uses repetition as a teaching strategy, while the learner model simulates a native Chinese speaker at an early stage of L2 English learning. Joint simulation may allow valuable insights into the entire learning process. In this study, speakers from the speechocean762 corpus were enlisted, using a word list that includes phonemes known to pose difficulties for Chinese speakers. The similarity between the output of the learning process and real learner data is evaluated using MCD, PPG, and wav2vec 2.0 distortion measures. The results indicate that the similarity between the process output and real learners with low proficiency is higher compared to that with real learners with high proficiency.

Index Terms— Second language acquisition, Teacher-learner interaction, Simulation, Pronunciation learning

1. INTRODUCTION

Language serves as a fundamental medium of human communication, enabling individuals to convey ideas, knowledge, and emotions. The growing prevalence of multilingualism, with approximately half of the world's population being multilingual [1], further emphasizes the increasing importance of research in the field of second language (L2) learning. All L2 learners are generally assumed to progress through distinct stages in their language acquisition journey, albeit at different rates. These stages encompass the pre-production stage to advanced fluency, with the early learning stage being particularly noteworthy [2]. Several studies have demonstrated the impact of first langugae (L1) interpretation and the challenges it presents for L2 learners. For instance, when Japanese speakers learn English as a second language, they may encounter difficulties with minimal pairs such as ”rocket” and ”locket” [3]. Effective teaching strategies, such as incorporating pronunciation-focused activities like practicing with minimal pairs, utilizing phonetic transcription systems, and encouraging learners to mimic native speakers, enable L2 learners to enhance their pronunciation skills, accelerate their language acquisition process, and achieve better proficiency outcomes [4]. According to Larsen-Freeman (2012), repetition is an effective teaching strategy [5] which involves the deliberate practice of repeating sounds, words, or phrases to improve pronunciation accuracy and fluency.

As mentioned earlier, the process of L2 learning is complex, influenced by factors such as the impact of the L1, teaching strategies, and learner behavior. Simulation, in this context, refers to employing computational models to mimic and simulate the L2 learning process. By simulating and examining the interplay of these factors, a clearer understanding of their relationships can be achieved, leading to informed decisions on how to optimize the process of L2 learning [6]. This optimization can be investigated by observing the progression of learner proficiency over time.

A speech corpus provides invaluable training data for computational models in language learning. The widely recognized TIMIT corpus [7] stands as a notable example, extensively employed for studying speech by including data from native English speakers across various dialects and languages. Furthermore, other publicly available corpora, such as L2-ARCTIC, a non-native English speech corpus with manual annotations [8], and the Sell-corpus, a Chinese-English speech corpus [9], cater specifically to pronunciation assessment. These corpora, equipped with phoneme-level annotations, play a pivotal role in studying pronunciation accuracy within the realm of language learning [10].

Evaluation metrics and performance measurement are essential for assessing the effectiveness and quality of language learning systems and models [11]. These metrics serve as objective measures to evaluate learners' performance, track their progress, and assess the impact of instructional approaches. Pitch Periodicity Glottal-Derivative Dynamic Time Warping (PPG-DTW) is the method employed in this study [12] to analyze the data collected during the 42-day Shadowing Marathon training. This method enables the examination of gradual changes in both L2 perception and production. Goodness of Pronunciation (GOP) is a specific objective approach for evaluating learners' pronunciation. By analyzing factors such as phonetic accuracy, stress, intonation, and rhythm, GOP offers valuable insights into the overall

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

pronunciation abilities of learners [13]. In addition to GOP, another approach for language assessment is the Mispronunciation Detection and Diagnosis (MDD) model. Traditional language models often overlook mispronunciations, thus requiring robust acoustic modeling to differentiate between native productions with canonical phonetic pronunciations and non-native pronunciations [14].

Intelligent Tutoring Systems (ITS) play a significant role in modern education by simulating and augmenting the teacher role, providing personalized instruction, feedback, and support to students [15]. By passively offering personalized feedback, ITS analyze learners' performance, identify areas for improvement. Through the use of computational models, ITS create interactive and tailored learning experiences for individualized language acquisition. The progress in computational models, specifically in natural language processing, user modeling, and intelligent tutoring systems (ITS), has driven the expansion of the discipline referred to as Computer-Assisted Language Learning (CALL) [16]. CALL utilizes computational tools, interactive software applications, and online platforms to offer learners personalized instruction, prompt feedback, and interactive language practice activities.

Computational models of L2 learners consider both their perception and production of speech sounds. L2 speech perception models, such as the Perceptual Assimilation Model (PAM), are frameworks that aim to explain how L2 learners perceive and categorize speech sounds [17]. PAM, in particular, focuses on how learners assimilate L2 speech sounds into their existing phonological system. PAM proposes that L2 speech sounds are categorized based on their perceived similarity to the sounds of the learners' L1. The speech learning model (SLM-r) is a computational model that aims to explain the process of speech perception and production in L2 learners [18]. SLM-r proposes that L2 speech learning involves a feedback loop between perception and production. Through practice and production, learners gradually refine their production of speech sounds, aligning them with their target language's phonological system.

In the field of L2 learning, computational models exist for either the teacher or the learner, but none have effectively integrated both roles. The objective of this study is to simulate the interaction between teacher and learner in the English pronunciation learning process using deep learning algorithms. Simulating this interaction is essential for exploring and experimenting with different teaching and learning strategies. Furthermore, simulation allows for the manipulation of various variables and conditions that may be impractical to explore in real-world settings.

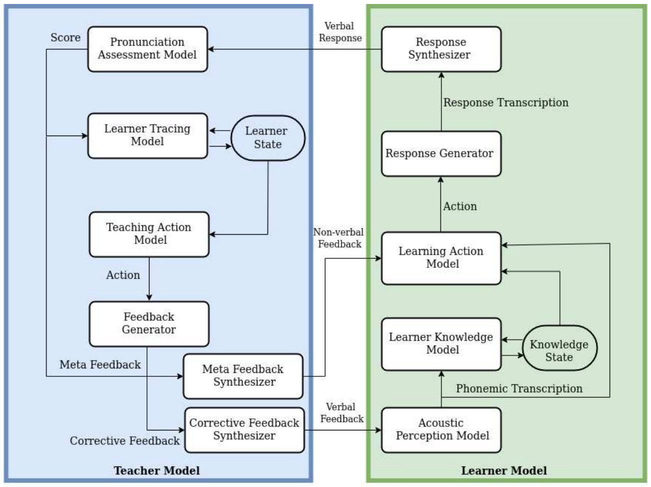

Fig. 1. The general proposed model for teacher-learner interaction in English language pronunciation learning.

2. SIMULATION OF TEACHER-LEARNER INTERACTION

This section focuses on the development of the proposed model of simulation of teacher-learner interaction in English language pronunciation learning. The development of the learner model depicted in Figure 1, which captures how learners perceive, recognize, and link speech to the production model, is based on the contributions of researchers in the field [17, 18]. These studies have provided valuable insights into the cognitive processes involved in language learning. Similarly, the design of the teacher model in the proposed system, as shown in Figure 1, has been informed by the findings of studies on teacher modeling [16, 15]. By building upon these previous works, our proposed model in Figure 1 integrates both the learner and teacher models to simulate the interaction between them within the context of English pronunciation learning as L2. Section 2.1 provides further details on the teacher model, with its implementation discussed in Section 3.1. Similarly, Section 2.2 offers comprehensive information on the learner model, accompanied by its corresponding implementation details in Section 3.2.

2.1. Teacher model

The teacher model in the proposed system simulates an English native speaker and employs repetition as a teaching strategy, where words are repeated multiple times with varying pronunciations. The repetition continues until the learner model achieves satisfactory pronunciation. This indicates the completion of the learning process within the proposed model. The integrated pronunciation assessment model analyzes the learner's phonemes and the output is used as scores for each perceived phoneme from the learner model. The learner tracing model updates the learner state based on the score and current state. The teaching action model utilizes the learner state to determine appropriate instructional actions, guided by a learner state. The feedback generator model produces corrective feedback in the form of a sequence of phonemes. This phoneme sequence is then converted into speech representation and synthesized to deliver customized corrective feedback. These models work together to provide a corrective feedback, enabling learner model to improve their pronunciation effectively.

2.2. Learner model

The learner model within the proposed system simulates a Chinese learner in the early stages of English learning. The learner model actively engages by repeating after the teacher model until the completion of the learning process, as determined by the teacher model. The perception model in the learner model receives verbal feedback, which is a repetition of the intended word, and recognizes it as a sequence of phonemes. The knowledge state, representing the acquired phonemes recognized by the acoustic perception model, is then updated by the learner knowledge model. The learning action model utilizes the knowledge state and the perceived phonemic transcription to determine the next learning action. This model operates on the concept that learning behavior and actions can be modeled as a sequence of decisions, informed by the response generator model. The generated response text is sent to an acoustic feature generator, which converts it into a speech representation. Finally, the response synthesizer synthesizes the speech representation to the teacher model. These models collaborate to enhance learner model pronunciation, collectively providing a more refined response to the teacher model's feedback.

3. EXPERIMENTAL SETUP

3.1. Teacher Model Implementation

The pronunciation assessment model, implemented using GOP approach [13] with the Kaldi tool [19] and the WSJ-CAM0 British English corpus [20], using a training set of 15.5 hours of speech from 92 speakers, a development set of 2.25 hours from 18 speakers, and a test set from 48 speakers. The corpus provides detailed transcriptions for all the utterances. The output of the pronunciation assessment model is used as scores for each perceived phoneme from the learner model. The learner tracing model updates the learner state based on a rule-based approach. If the score from the pronunciation assessment model matches the average score of the teacher model feedback, the learner state transitions to the final state. Otherwise, the learner state remains in the learning state. The teaching action model determines the next action using a rule-based model. If the learner state is in the final state, the teaching process stops; otherwise, it continues. The feedback generator model, also employing a rule-based model, generates subsequent feedback based on the selected action. In cases where further teaching is required, the phoneme sequence is passed to the next model. For the generation of corrective feedback, the system employs Fastspeech2 [21]. This model has been trained on the LJ speech corpus [22], a publicly available English speech corpus consisting of 13,100 short audio recordings from a single speaker.

3.2. Learner Model Implementation

The acoustic perception model outputs the phonemic transcription of the perceived feedback received from the teacher model. The learner model simulates the English language learning process for a Mandarin native speaker. This involves refining both the Mandarin acoustic model and language model within the acoustic perception model. To prepare the data, the AISHELL-1 corpus [23] was used to train an acoustic model with Kaldi tool [19]. Forced alignment was then performed on the audio files accompanied by word-level transcripts, and the phoneme-level transcript was extracted. The training and testing corpus were then prepared using the phone-level transcript, and features were extracted using high-resolution Mel-frequency cepstral coefficients. The learner knowledge model updates the knowledge state, which represents the most frequently perceived phonemic transcription. The learning action model determines the next action using a rule-based approach. Each perceived phonemic transcription is compared with the knowledge state, and if the nonverbal feedback increases, the action is to respond with the perceived phonemic transcription. Otherwise, the action is to choose the most recent phonemic transcription from the knowledge state. Finally, the response synthesizer, implemented using the Fastspeech2 model [21] trained on the AISHELL-3 corpus [24], is utilized to synthesize a verbal response from the generated phoneme sequence. The AISHELL-3 corpus is a multi-speaker Mandarin audio corpus containing a total of 88,035 recordings from 218 native speakers.

3.3. Corpus processing

A reference set was prepared using the speechocean762 corpus [25] to evaluate the similarity between the output of the proposed learning process described in Section 2. This corpus, which consists of 5,000 English utterances collected from 250 Mandarin speakers, is freely available as an open-source resource. Each audio file is assigned five attribute scores at the utterance level, ranging from 0 to 10. These scores assess the accuracy, fluency, completeness, prosody, and overall quality of the audio. The evaluation process involves five expert evaluators, who independently score each utterance according to the same metrics. The reported score is the average of their assessments. The training set comprises

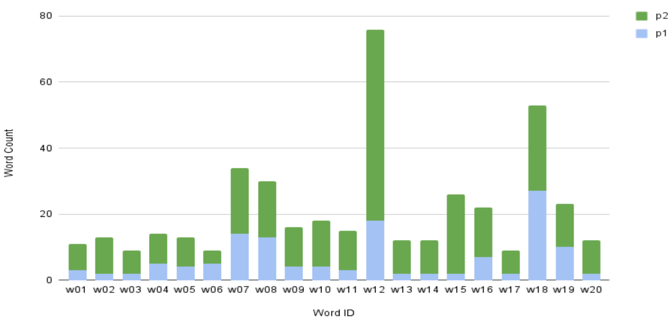

Fig. 2. A list of 20 words with two different proficiency levels. p1 and p2 denotes low and high proficiency respectively

2,500 utterances, 15,849 words, and 47,076 phones, while the testing set consists of 2,500 utterances, 15,967 words, and 47,369 phones. These sets are combined and used as the reference set.

Proficiency calculation The proficiency of each speaker was assessed by calculating the average reported accuracy score for all their utterances. In order to investigate the influence of language proficiency on L2 learning, the study categorized the speakers in the reference set into two groups: low proficiency (p1) and high proficiency (p2). Table 1 provides details on the proficiency range, number of speakers, and word count for p1 and p2.

Table 1. Proficiency range, Number of speakers, Number of utterance and number of words in p1 and p2

	Proficiency range	#Speakers	#Uttrenc	#Words
p1	3.95 - 7.5	142	2,840	16,960
p2	7.5 - 9.55	43	860	4,105

To simulate the early stages of language learning, a learner model was employed, which utilized an acoustic perception model and a response synthesizer trained in Mandarin. Consequently, the output of the learner model's utterances should exhibit more similarities with the utterances of p1 speakers, as compared to the output generated by the learner model and the utterances of p2 speakers.

Word-list Selection The word list selection process was guided by previous research, specifically focusing on identifying phonemes that Chinese speakers commonly struggle with when learning English [26]. A meticulously curated word list of 20 words was then assembled, with each word carefully selected to specifically address one or more of these challenging phonemes. The primary goal was to highlight the distinctions between the phonemes in L1 and L2, enabling a more precise analysis of their pronunciation abilities. Additionally, a random word ID was assigned to each word in the list. Figure 2 shows the visual representation of the word count in both p1 and p2.

Speakers Selection To ensure the quality of the speech data and address potential issues related to children's speech [27], a sample of speakers aged 10 years and above was specifically chosen. A total of 175 speakers are included in this study, encompassing a diverse mix of male and female participants. For each word, two random speakers from p1 and two random speakers from p2 were selected, allowing for a well-rounded representation across proficiency groups.

4. RESULTS AND DISCUSSION

4.1. Evaluation metrics

The system's performance evaluation includes various metrics. Among them, mel-cepstral distortion (MCD) measures the difference between sets of mel-cepstral coefficients, representing the speech signal's spectral envelope [28]. Additionally, distortion metrics based on wav2vec 2.0 features and phonetic posteriorgrams are assessed. Wav2vec 2.0 features capture high-level representations of speech signals [29], while phonetic posteriorgrams provide information about the distribution of phonetic content [30]. These metrics comprehensively evaluate aspects such as spectral fidelity, phonetic accuracy, and other relevant features of the generated or modified speech signals.

4.2. Mel cepstral distortion

The MCD measures spectral distortion between some source and target mel cepstral coefficients. It is computed using the following equation, The applied MCD, between the output of the proposed learning process and reference set, offers a robust quantitative measure for assessing the perceptual differences between two sets of mel-cepstral coefficients

$$\text{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{i=1}^{k}(MCC_i^c - MCC_i^t)^2} \qquad (1)$$

where, i represent mel cepstral coefficients index, MCC_i^c and MCC_i^t denote i^{th} dimensional coefficient of the converted and target coefficients, respectively. In Figure 3, the mean MCD values obtained from 100 samples per word are reported. For each of the word, two MCD values are computed: first between learner model output and low proficient speakers (p1) from speechocean762 corpus and second between learner model output and high proficient speakers (p2) from speechocean762 corpus. From Figure 3 it is observed that 60% of the samples have lower MCD values, 20% have similar MCD and 20% have higher MCD values.

4.3. Phonetic posteriorGrams distortion

The Phonetic PosteriorGrams (PPG) representation is designed to capture speech characteristics while suppressing extralinguistic factors [30]. It achieves this by converting a

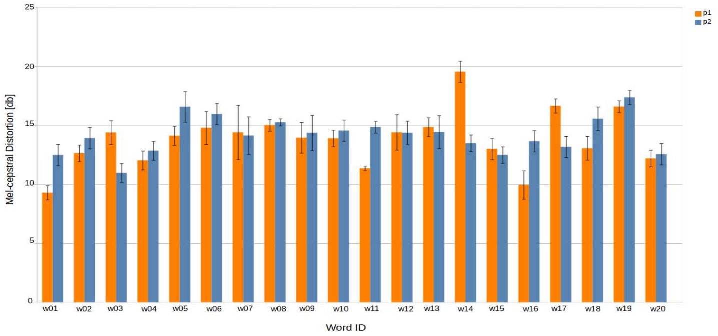

Fig. 3. Plot of MCD scores obtained for 20 words. p1 and p2 denotes low and high proficiency respectively

speech frame into a posterior probability distribution over a comprehensive set of context-dependent phones. In Figure 4, the PPG-based distortion is presented for two proficiency levels. For each word, two PPG distortion values are computed: the first between the learner model output and p1 speakers, and the second between the learner model output and p2 speakers. Figure 4 demonstrates that the majority of words with low proficiency exhibit lower PPG distortion.

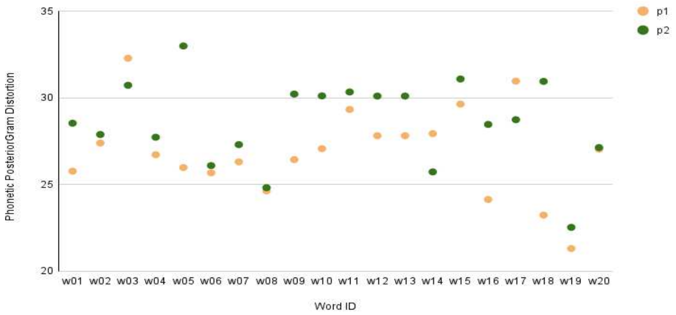

Fig. 4. Plot of PPG based distortion obtained for 20 words. p1 and p2 denotes low and high proficiency respectively

4.4. Wav2vec 2.0 distortion

In addition to MCD and PPG-based distortions, wav2vec 2.0-based distortions were also analyzed [29]. A pretrained wav2vec 2.0 model was utilized to extract frame-level representations. These representations capture both the acoustic and linguistic properties of the audio [31]. Similar to MCD and PPGs, the wav2vec 2.0-based distortion is computed between the learner model output and the corresponding word sample from the speechocean762 corpus. The wav2vec 2.0-based distortion is obtained by first aligning the features from two groups using DTW. Subsequently, the euclidean distance is computed for the aligned features, and the mean of all values across all frames is calculated. In Figure 5, the wav2vec 2.0-based distortion is shown for two proficiency levels. For each word, two wav2vec 2.0-based distortion values are computed: first between the learner model output and low proficient speakers (p1), and second between the learner model output and high proficient speakers (p2). Figure 5 depicts that the wav2vec 2.0-based distortion is lower for all low proficiency speakers.

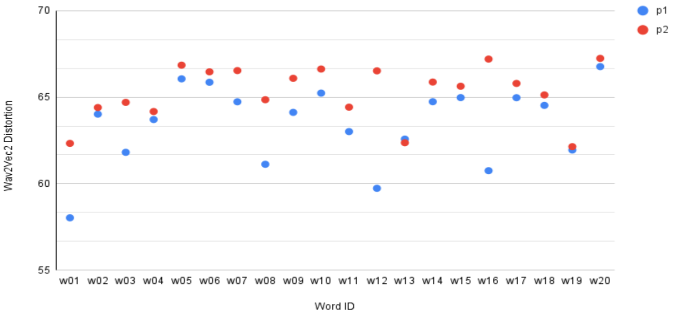

Fig. 5. Plot of wav2vec 2.0 based distortion obtained for 20 words. p1 and p2 denotes low and high proficiency respectively

5. CONCLUSIONS

This paper presents a computational model that captures the teacher-learner interaction in L2 learning, providing valuable insights into the learning process. Through joint simulation of the teacher and learner models, a comprehensive evaluation was conducted, comparing the model's output to real learner data from the speechocean762 corpus, considering both low and high proficiency levels. The evaluation encompassed multiple metrics, including Mel cepstral distortion (MCD), which revealed that 60% of the samples had lower values, 20% had similar values, and 20% had higher values, indicating the model's effectiveness in spectral precision. Additionally, the evaluation encompassed wav2vec 2.0 features and phonetic posteriorgrams, showcasing lower PPG-based distortion for most low-proficiency words, while wav2vec 2.0 distortion was lower across the board. In essence, the lower distortion suggests that the computational model is successfully simulating the pronunciation challenges faced by learners at an early stage of L2 English learning, making it a valuable tool for understanding and potentially improving the L2 learning process.

6. ACKNOWLEDGEMENT

This work was partially supported by Taif University in Saudi Arabia and the Royal Embassy of Saudi Arabia Cultural Bureau in the United Kingdom.

7. REFERENCES

[1] T. Tokuhama-Espinosa, *The Multilingual Mind: Issues Discussed By, For, and about People Living with Many Languages*. Greenwood Publishing Group, 2003.

[2] C. Perdue, “Pre-basic varieties: The first stages of second language acquisition,” *Toegepaste taalwetenschap in artikelen*, vol. 55, no. 1, pp. 135–149, 1996.

[3] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, “Pperceived phonetic dissimilarity and l2 speech learning: the case of japanese /r/ and english /l/ and /r/,” *Journal of Phonetics*, vol. 32, no. 2, pp. 233–250, 2004.

[4] A. P. Gilakjani, “A study of factors affecting efl learners’ english pronunciation learning and the strategies for instruction,” *International journal of humanities and social science*, vol. 2, no. 3, pp. 119–128, 2012.

[5] D. Larsen-Freeman, “On the roles of repetition in language teaching and learning,” *Applied Linguistics Review*, vol. 3, no. 2, pp. 195–210, 2012.

[6] B. MacWhinney, “Computational models of child language learning: an introduction,” *Journal of Child language*, vol. 37, no. 3, pp. 477–485, 2010.

[7] L. D. Consortium *et al.*, “The darpa timit acoustic-phonetic continuous speech corpus,” *NIST Speech CD*, pp. 1–1, 1990.

[8] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-arctic: A non-native english speech corpus.,” in *Interspeech*, pp. 2783–2787, 2018.

[9] Y. Chen, J. Hu, and X. Zhang, “Sell-corpus: an open source multiple accented chinese-english speech corpus for l2 english learning assessment,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7425–7429, IEEE, 2019.

[10] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, “An end-to-end mispronunciation detection system for l2 english speech leveraging novel anti-phone modeling,” *arXiv preprint arXiv:2005.11950*, 2020.

[11] H. Dalianis and H. Dalianis, “Evaluation metrics and evaluation,” *Clinical Text Mining: secondary use of electronic patient records*, pp. 45–53, 2018.

[12] T. Kunihara, C. Zhu, N. Minematsu, and N. Nakanishi, “Gradual improvements ob-served in learners’ perception and production of l2 sounds through continuing shadowing practices on a daily basis,” in *Proc. Interspeech*, 2022.

[13] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, “An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities.,” in *INTERSPEECH*, pp. 954–958, 2019.

[14] M. Wu, K. Li, W.-K. Leung, and H. Meng, “Transformer based end-to-end mispronunciation detection and diagnosis.,” in *Interspeech*, pp. 3954–3958, 2021.

[15] V. Slavuj, B. Kovačić, and I. Jugo, “Intelligent tutoring systems for language learning,” in *2015 38th MIPRO*, pp. 814–819, IEEE, 2015.

[16] R. Shadiev and M. Yang, “Review of studies on technology-enhanced language learning and teaching,” *Sustainability*, vol. 12, no. 2, p. 524, 2020.

[17] O.-S. Bohn and M. J. Munro, *Language experience in second language speech learning: In honor of James Emil Flege*, vol. 17. John Benjamins Publishing, 2007.

[18] J. E. Flege and O.-S. Bohn, “The revised speech learning model (slm-r),” *Second language speech learning: Theoretical and empirical progress*, pp. 3–83, 2021.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF, IEEE Signal Processing Society, 2011.

[20] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition,” in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 81–84, IEEE, 1995.

[21] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.

[22] K. Ito, “The lj speech dataset.” https://keithito.com/LJ-Speech-Dataset/, 2017.

[23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “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)*, pp. 1–5, IEEE, 2017.

[24] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “Aishell-3: A multi-speaker mandarin tts corpus and the baselines,” *arXiv preprint arXiv:2010.11567*, 2020.

[25] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, “speechocean762: An open-source non-native english speech corpus for pronunciation assessment,” *arXiv preprint arXiv:2104.01378*, 2021.

[26] S. Khanal, M. T. Johnson, M. Soleymanpour, and N. Bozorg, “Mispronunciation detection and diagnosis for mandarin accented english speech,” in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 62–67, IEEE, 2021.

[27] A. Potamianos and S. Narayanan, “Robust recognition of children’s speech,” *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[28] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1, pp. 125–128, IEEE, 1993.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[30] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, “Personalized, cross-lingual tts using phonetic posteriorgrams.,” in *Interspeech*, pp. 322–326, 2016.

[31] E. Islam, C. Park, and T. Hain, “Exploring speech representations for proficiency assessment in language learning,” in *9th Workshop on Speech and Language Technology in Education (SLaTE) Proceedings*, pp. 151–155, International Speech Communication Association (ISCA), 2023.